Volume 24 – Number 1 – 2009 – ISSN 0175-1336

JLCL

Journal for Language Technology
and Computational Linguistics

Automatic Genre Identification: Issues and Prospects

Edited by
Marina Santini, Georg Rehm, Serge Sharoff
and Alexander Mehler

GSCL Gesellschaft für Sprachtechnologie und Computerlinguistik www.gscl.info

Automatic Genre Identification: Issues and Prospects

Impressum

JLCL
ISSN 0175-1336
Volume 24 (1) – 2009

Journal for Language Technology and Computational Linguistics – offizielles Organ der GSCL

Herausgeber

Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Prof. Dr. Alexander Mehler, Universität Bielefeld
alexander.mehler@uni-bielefeld.de
Prof. Dr. Christian Wolff, Universität Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Anschrift der Redaktion

Prof. Dr. Christian Wolff
Universität Regensburg
Institut für Medien-, Informations- und Kulturwissenschaft
D-93040 Regensburg

Wissenschaftlicher Beirat

Vorstand, Beirat und Arbeitskreisleiter der GSCL
http://www.gscl.info/vorstand.html
http://www.gscl.info/

Erscheinungsweise

2 Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober. Preprints und redaktionelle Planungen sind über die Website der GSCL einsehbar (*http://www.gscl.info*).

Einreichung von Beiträgen

Eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall elektronisch und zusätzlich auf Papier übermittelt werden. Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind an die Herausgeber zu übermitteln.

Bezugsbedingungen

Für Mitglieder der GSCL ist der Bezugspreis des JLCL im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von 25,- € (inkl. Versand), Einzelexemplare zum Preis von 15,- € (zzgl. Versandkosten) bei der Redaktion bestellt werden.

Satz und Druck

Dietmar Esch, Christian Groszewski, Bielefeld,
mit *LaTeX* (*pdfeTeX* / *MiKTeX*) und *Adobe InDesign CS*
Druck: Druck TEAM KG, Regensburg

Marina Santini, Georg Rehm, Serge Sharoff, Alexander Mehler

Editorial

In recent years, a multitude of most interesting research has been carried out in linguistics, psycholinguistics, computational linguistics and information retrieval with regard to the topic of web genres. Despite the increasing interest in this novel and innovative field within different communities, there is still a significant lack in literature, especially concerning edited collections and journal issues that provide an overview of recent research. The aim of this special issue of the *Journal for Language Technology and Computational Linguistics* is to contribute to filling this gap. More specifically, this issue is dedicated to automatic genre identification.[1]

Genres are categories that subsume texts which have multiple features in common, most importantly, a shared communicative purpose. Genres form and evolve within specific discourse communities, are instantiated as well as enforced and most often also given a name by members of their respective discourse communities. An important characteristic is that their users are able to recognise certain genres, such as, for example, an invoice, a business letter, a shopping list, or a menu very quickly based on genre-specific properties (such as, for example, a conventionalized text structure, certain key words, a specific layout and several other properties). Recognizing that a text belongs to a certain genre helps in the assessment of its importance and significance as well as the communicative goals its original author associated with the text. In other words, genres are a very important means for effective communication.

Karlgren and Cutting (1994) and Kessler et al. (1997) were among the first to suggest that genres might be useful to enhance information retrieval systems. The automatic identification of text genres could be used to generate additional metadata about text collections that, in turn, could enable users to search for texts based on their genre properties (with hypothetical queries such as "find instances of the genres *invoice* or *business letter* that contain *company car*"). Later on, when it became evident that computational linguistics had a new, exciting and most challenging target in the form of the World Wide Web, this idea was extended to web documents (see, for example, Crowston and Williams, 1997; Haas and Grams, 2000; Roussinov et al., 2001; Rehm, 2002). If we could identify automatically the genres of web documents we could extend the procedure and functionality sketched above to the whole web, giving users new and innovative means of locating relevant content online.

There is a plethora of open questions with regard to genre and web genre research. People who approach the field from a linguistic and text linguistic perspective try to find ways of describing and representing genres and web genres with the help of knowledge representation formalisms such as ontologies. Another interesting question concerns the evolution of genres and web genres, how they are formed in dynamic social communication system and discourse communities and how certain optional properties

[1] It is interesting to note that most articles in this issue are authored or co-authored by students who are still engaged in ongoing research or who have recently completed their research projects.

or features of genre instances slowly change their status into obligatory components. Researchers who work in psycholinguistics are, for example, concerned with the problem of genre recognition: how do we identify instances of certain genres, what kind of cues or inherent properties of a document do we employ in order to categorize a specific text into a certain genre? Is it primarily words that we use for this process or layout features or probably a significant document structure? Do ordinary web users even think in terms of "genres", do they categorize web documents into different types? Computational linguists try to fit all of these currently still fragmented pieces and insights together in order to build systems that are able to identify genres and especially web genres automatically. A few of these open questions are of utmost importance: how do web genres work in the hypertext environment of the World Wide Web with regard to the document, sub-document and super-document level? What kind of features could or should be used and extracted from proper documents in order to compute their respective web genres? Can scalable systems be built that are able to identify hundreds of different genres – both traditional genres and genuine web genres? How can a reference corpus of web genres be built so that the underlying theoretical assumptions inherently encoded into the corpus are met by the probably differing theoretical opinions of other researchers?

This special issue and other activities its editors and contributors are involved in has its origin in 2007. Intense networking among members of the genre community lead to a number of events. More precisely, in July 2007, Marina Santini and Serge Sharoff organised the colloquium "Towards a Reference Corpus of Web Genres" (Santini and Sharoff, 2007) which was held at Corpus Linguistics 2007 in Birmingham, UK.[2] Shortly afterwards, Marina Santini and Georg Rehm organised a follow-up workshop, "Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing" (Rehm and Santini, 2007) which was held in conjunction with the conference Recent Advances in Natural Language Processing (RANLP) in Borovets, Bulgaria, in September 2007.[3] One result of this workshop was a common paper on the construction of a reference corpus of web genres (Rehm et al., 2008). Yet another event organised in 2007 was the panel[4] arranged by Mark Rosso, "Towards the Use of Genre to Improve Search in Digital Libraries: Where Do We Go from Here?". This panel was sponsored by SIG-CR and SIG-HCI and held in conjunction with ASIST 2007 in Milwaukee, Wisconsin (USA). It lead to the publication of the special issue "Bringing genre into focus" (Freund and Ringlstetter, 2008). At roughly the same time, the *Scandinavian Journal of Information Systems* published a special issue titled "Genre Lens on Information Systems" (Päivärinta et al., 2008) which brings into focus an additional perspective on genre usefulness.[5]

The interest in genre research has not declined over the years, the year 2009 is also rich of genre-related activities. First, the workshop "Automated Document Genre Classification Workshop: Supporting Digital Curation, Information Retrieval, and

[2] http://corpus.leeds.ac.uk/serge/webgenres/colloquium/
[3] http://www.sics.se/use/genre-ws/
[4] http://www.asis.org/Conferences/AM07/panels/18.html
[5] http://iris.cs.aau.dk/index.php/volume-20-40200841-no-1.html

Knowledge Extraction"[6] continues the general discussion and tries to establish a common roadmap for future projects. Second, the forthcoming book *Genres on the Web: Computational Models and Empirical studies* (Mehler et al., Forthcoming)[7] presents a wide range of conceptualisations of the notion of web genre together with an overview of computational approaches to web genre classification and structure modelling. Finally, this special issue of the *Journal for Language Technology and Computational Linguistics*, which has been conceived as the book's companion volume, is now available online with a collection of recent genre research. Taken together, the book and this JLCL special issue show the most up to date and comprehensive range of theoretical, computational and empirical research on web genres.

Structure of this Special Issue

Research on the automatic identification of web genres usually falls in one of three different categories, i. e., theoretical approaches, implementations, and the creation of resources for analysis and evaluation.

Theoretical approaches describe the extremely hard problem that web genres are able both to span multiple documents (i. e., a hypertext) as well as to instantiate only a small part of a single web page. There are, however, evolutionary processes at work that are responsible for the existence of multiple conventions and preferred instances of genres and web genres (see, for example, Rehm, 2007; Mehler, 2009). Due to these evolutionary processes, web genres come into being and, afterwards, slowly and continuously change. In their article "The Evolution of Genre in Wikipedia", Malcolm Clark, Ian Ruthven and Patrik O'Brian Holt examine how genres develop over a period of six years in the online encyclopedia Wikipedia. The authors concentrate on the biographical article and generate a number of follow-up research questions as well as plans for experimental work. A related area of research is addressed by Philip M. McCarthy, John C. Myers, Stephen W. Briner, Arthur C. Graesser and Danielle S. McNamara who present "A Psychological and Computational Study of Sub-Sentential Genre Recognition". In their article the authors describe three experiments on genre recognition based on words alone. Their conclusions not only have implications for research on the automatic identification of web genres but also for a better understanding of genre recognition in general.

Implementations are concerned with the creation of actual systems that accomplish the overall goal of this field of research. When building a genre identification system, the decision of which features to use for this process is very important. In their paper "Cost-Sensitive Feature Extraction and Selection in Genre Classification", Ryan Levering and Michal Cutler describe a complex approach for the automatic selection of features for the task of genre identification and report experimental results on two datasets. Chaker Jebari concentrates on processes that are able to categorize documents into multiple genres. In his paper "A New Centroid-based Approach for Genre Categorization of Web Pages", Jebari uses machine learning algorithms in order to compute multiple

[6] http://www.dcc.ac.uk/events/genre-classification-2009/

[7] http://sirao.kgf.uni-frankfurt.de/webgenrebook/index.html

ranks for each documents. This approach reflects the fact that a single web page often contains instances of multiple genres. It is this core problem of web genre research that is also addressed by the article "Multi-Label Approaches to Web Genre Identification". In their paper, Vedrana Vidulin, Mitja Luštrek and Matjaž Gams extract multiple features from web pages in order to test the performance of several classifiers for the task of assigning multiple genre labels to a document.

The creation of resources is closely related to the two categories mentioned above and deals with web genre corpora and datasets. In their article "Building a Corpus of Italian Web Forums: Standard Encoding Issues and Linguistic Features", Silvia Petri and Mirko Tavosanis examine linguistic properties of postings in web discussion groups and construct a corpus of these documents. They use an encoding and annotation scheme that is based on the TEI guidelines. Undoubtedly, one of the most significant gaps in current web genre research is the lack of a reference corpus of web genres that interested parties could use to evaluate their own systems based on a shared resource that was built specifically for evaluation purposes (see, for example, Rehm et al., 2008). The article "Web Genre Benchmark Under Construction" by Marina Santini and Serge Sharoff discusses this problem in detail and suggests a solution.

Acknowledgments

This special issue would not have been possible without the time and dedication of the following reviewers: Eric Atwell, Malcom Clark, Luanne Freund, Mikael Gunnarsson, Theresa Heyd, Marie-Paule Jacques, Yunhyong Kim, Nedim Lipka, Mark Rosso and Efstathios Stamatatos. Further, we gratefully acknowledge support of the German Federal Ministry of Education (BMBF) through the research project *Linguistic Networks* and of the German Research Foundation (DFG) through the Excellence Cluster 277 *Cognitive Interaction Technology* via the Project *Knowledge Enhanced Embodied Cognitive Interaction Technologies* (KnowCIT), and the Research Group 437 *Text Technological Information Modeling* via the Project A4 *Induction of Document Grammars for Webgenre Representation*.

References

Crowston, Kevin and Williams, Marie (1997): "Reproduced and Emergent Genres of Communication on the World-Wide Web". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. IEEE Computer Society, volume 6, pp. 30–39.

Freund, Luanne and Ringlstetter, Christoph (2008): "Special Issue: Bringing Genre into Focus". *Bulletin of the American Society for Information Science and Technology* 34 (5).

Haas, Stephanie W. and Grams, Erika S. (2000): "Readers, Authors, and Page Structure – A Discussion of Four Questions Arising from a Content Analysis of Web Pages". *Journal of the American Society for Information Science* 51 (2): pp. 181–192.

Karlgren, Jussi and Cutting, Douglass (1994): "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". In: *COLING 94 – The 15th International Conference on*

Computational Linguistics. Association for Computational Linguistics, Kyoto, volume 2, pp. 1071–1075.

Kessler, Brett; Nunberg, Geoffrey and Schütze, Hinrich (1997): "Automatic Detection of Text Genre". In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann, pp. 32–38.

Mehler, Alexander (2009): "A quantitative graph model of social ontologies by example of Wikipedia". In: *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, edited by Dehmer, Matthias; Emmert-Streib, Frank and Mehler, Alexander, Boston/Basel: Birkhäuser.

Mehler, Alexander; Sharoff, Serge and Santini, Marina (editors) (Forthcoming): *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Päivärinta, T.; Shepherd, M.; Svensson, L. and Rossi, M. (2008): "Special Issue Genre Lens on Information Systems". *Scandinavian Journal of Information Systems* 20 (1).

Rehm, Georg (2002): "Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii: IEEE Computer Society.

Rehm, Georg (2007): *Hypertextsorten: Definition – Struktur – Klassifikation*. Norderstedt: Books on Demand. (PhD thesis in Applied and Computational Linguistics, Giessen University, 2005).

Rehm, Georg and Santini, Marina (editors) (2007): *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, Borovets, Bulgaria. Held in conjunction with RANLP 2007.

Rehm, Georg; Santini, Marina; Mehler, Alexander; Braslavski, Pavel; Gleim, Rüdiger; Stubbe, Andrea; Symonenko, Svetlana; Tavosanis, Mirko and Vidulin, Vedrana (2008): "Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems". In: *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.

Roussinov, Dmitri; Crowston, Kevin; Nilan, Mike; Kwasnik, Barbara; Cai, Jin and Liu, Xiaoyong (2001): "Genre Based Navigation on the Web". In: *Proceedings of the 34th Hawaii International Conference on Systems Sciences (HICSS-34)*. IEEE Computer Society.

Santini, Marina and Sharoff, Serge (editors) (2007): *Proceedings of the Colloquium Towards a Reference Corpus of Web Genres*, Birmingham, UK. Held in conjunction with Corpus Linguistics 2007.

Table of Contents

JLCL

Journal for Language Technology and Computational Linguistics – Volume 24(1) – 2009
Automatic Genre Identification: Issues and Prospects

Malcolm Clark, Ian Ruthven and Patrik O'Brian Holt

The Evolution of Genre in Wikipedia

This paper presents an overview of the ways in which genres, or structural forms, develop in a community of practice, in this case, Wikipedia. Firstly, we collected data by performing a small search task in the Wikipedia search engine (powered by Lucene) to locate articles related to global car manufacturers, for example, British Leyland, Ferrari and General Motors. We also searched for typical biographical articles about notable people, such as Spike Milligan, Alex Ferguson, Nelson Mandela and Karl Marx. An examination of the data thus obtained revealed that these articles have particular forms and that some genres connect to each other and evolve, merge and overlap. We then looked at the ways in which the purpose and form of a biographical article have evolved over six years within this community. We concluded the work with a discussion on the usefulness of Wikipedia as a vehicle for such genre investigations. This small analysis has allowed us to start generating a number of detailed research questions as to how forms may act as descriptors of genre and to discuss plans for experimental work aimed at answering these questions.

1 Introduction

The research reported and discussed in this paper combines information retrieval (IR), cognitive science and genre, merging and utilizing these for one particular purpose: to analyze how texts are used in different contexts with the final goal of retrieving structured texts. The main goals of effective IR are the identification of users' information needs and the evaluation of the results by creating IR applications that can discern better matches between users' information needs and the available documents (Clark, 2005). According to Ingwersen and Jarvelin (2005), IR is divided into computer science lab experiments versus 'user-orientated' social studies. Our approach is concerned with the latter and forms part of a wider human context to examine the ways in which the framework of a community of practice (CoP) (Wenger, 2000) gives rise to standardized information forms. The evolution of genre is an important part of this research and this paper describes the results of a preliminary study on genre development in Wikipedia In the recent past, the IR community, such as the text retrieval conference (TREC), and more recently, the initiative for the evaluation of XML retrieval (INEX) (Lalmas et al., 2004) have started to understand the importance of (technologically) structured text retrieval but up to now have largely overlooked two important concepts: naturally occurring structures called *genres* and the human perceptual processes which are used to identify and employ them. Genre has been discussed for centuries, most notably by Plato and, of course, by Aristotle, in his work on substance and form. Of course,

there is much more substance to Aristotle's doctrines but for this work, we plan to look at the ways in which humans extract the form which determines the nature of the observed object, in this case, Wikipedia articles or Electronic mail (Clark et al., 2008).

Genre (or kind), is used to differentiate between differing types of texts (especially in classical literature studies) such as reports, novels, poems, memoranda and so on. Although the form and function issue is central to genre theory, some theorists focus on the style, function, form and/or content of genre to distinguish between the 'kinds'. The aim of our discussion and research is to investigate genres and, particularly, how they evolve within Wikipedia. The Wikipedia Encyclopaedia, which first appeared in 2001, is growing and evolving day by day and has articles in more than 250 languages. Currently, the English version alone consists of more than 2.5 million articles and has more than 8 million registered editors (Ehmann et al., 2008). Only a small amount of genre analysis research utilising Wikipedia has as yet been carried out, but as Emigh and Herring (2005) pointed out, Wikipedia can offer an extraordinary insight into how a community can democratically participate in creating forms or genres to show the meaning of an article. Further to this, the work carried out by Collins et al. (2001) showed how there tend to be socially constructed communicative behaviours, namely genres, which emerge to improve the efficiency of the activities in a CoP. The purpose of this paper is to describe an initial study of these behaviours and the evolution of some articles in Wikipedia (English version only), in which classic forms of genres are found, such as Biographies. Some other types of 'new' structured genres, mainly defined by form and content, are also continuously evolving in the Wikipedia community. The question also arises, however, as to whether Wikipedia editors interact, discuss, debate and jointly learn? Does the community consist of the vital characteristics of a CoP, namely, "The Domain, The Community and The Practice," described by Wenger (see Section 2.2)? Our questions for this initial feasibility study were:

1. Is Wikipedia, as a CoP, a suitable vehicle for demonstrating the evolution/development of genre?

2. Are Wikipedia articles consistently composed of a combination of purpose and form?

3. What are the constituent parts of the CoP in the Wikipedia domain?

4. How does a classic genre, such as Biography, evolve in this community? Are there any possible new genres?

Section 2 begins with an introduction to genre, ecologies and CoPs. The third section examines the methodology for this study, the presence of Wikipedia genres by showing the results of a small search of genres and by mapping the genres. In part 3.3 there is a case study to take a closer look at the ways in which a biographical article has evolved since 2001. The conclusions drawn from the research and the plans for future work are presented in section four.

2 Genre, Ecologies and Communities of Practice

2.1 Genre

Genres have been around as an idea for thousands of years. Early examples can be found in the context of Plato's "ideas, forms or reality" and Aristotle's "rhetoric and poetics" (Aristotle, 1984). Aristotle disregarded Plato's musings on 'reality': he considered that whatever was perceivable by the individual was reality. He believed that the entire visual array was made up of *substance* and, most importantly for this research, *form*. Form was knowable, "which specified the individual and which could be abstracted from the objects in a process of perception. External objects impinged upon the senses, and due to the power of reason, the mind was able to extract the essence (or form), which determined the nature of the observed thing." (Breure, 2001). Contemporary authors writing on genre have continued with this theme, for example, Dewdney et al. (2001) refer to Substance and Form in their work. In the seminal book, 'Genre and the New Rhetoric',describe two prominent schools of thought based in different hemispheres: The North American School (heavily influenced by Miller 1984) and The Sydney School (heavily influenced by Halliday 1973, Halliday 1978, Kress and Threadgold 1988 and Martin 1999). In spite of the intrinsic differences between the two schools, some similarities can also be observed: they both acknowledge the superiority of the social in understanding genres and the role of context; in addition, they highlight the value of community or social factors. However, they do differ in other respects. The Sydney School focuses on the textual features in terms of linguistic analysis that stresses the static characteristics and rigid qualities. In contrast, the North American School emphasises the dynamic nature of genres, with the cornerstone of the theory based on interplay and interaction, and in particular, on the intricate associations between context and text. Both of these schools have implications for this work: not only are the textual features vital, but also the interaction and interplay of genres. There are also many genres that are of a static or dynamic nature.

Any thorough literature review of works on genre will reveal a general lack of consensus on the question of finding an appropriate definition for genre because so many questions remain unanswered as to how genres function, overlap and interact with each other, which rules and patterns constitute a genre and how these characteristics are perceived. We argue that the backgrounds of researchers influence the way they define genre, as Kwasnic and Crowston (2005) point out, the researcher chooses the definition appropriate to the current investigation. That said, there are significant similarities between scholars: compare, for example, Berkenkotter and Huckin (1995) *Situatedness* and *Duality of structure* with Yates and Orlikowski's *Genres of Organizational Communication*(1992). As Kwasnic and Crowston (2005) explain, the many definitions of genre and lack of agreement are not due to slipshod attitudes or lack of effort, but are rather indicative of the diversity of genre.

As Breure (2001) states: in most contemporary genre analysis, content and form are supplemented by purpose and function. In the context of this paper, it is the set

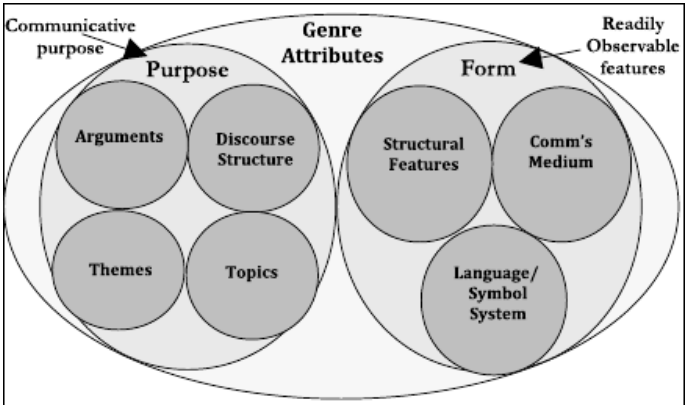


Figure 1: Orlikowski and Yates (1994) devised from the definition and attributes of purpose and form being used for this paper.

of structures and layout that show the user the documents' purpose (substance) and form through its structure, regardless of the topical nature of the writing. Figure 1 illustrates the definitions of purpose and form provided by Yates and Orlikowski (1992, 301-303), Orlikowski and Yates (1994, 544-545) and Yates et al. (1997, 550-551) that was influenced by Giddens (1986) structuration theory. Form, in the context of this project, simply refers to the easily perceptible features of the communication, such as those found in calls for papers, which include:

1. **Structural Features**: text formatting devices such as lists and headings, and devices for structuring interactions at meetings, such as agenda and chairpersons.

2. **Communication Medium**: pen and paper, telephone, or face to face.

3. **Language or Symbol System:** linguistic characteristics, such as the level of formality and the specialized vocabulary of corporate or professional jargon.

The purpose of the genre refers to the communicative purpose, in particular the social motives, themes and topical nature assembled and perceived in the communicative genre, for example, the purpose of a shareholders' meeting is to present the company's past accomplishments and future outlook to stockholders, or the purpose of a curriculum vitae is to summarise an individual's educational and employment history for a potential employer[1]. This particular technique defined by Orlikowski and Yates, and used by Emigh and Herring (2005) analysed genre by looking at the common and shared purpose to typical aspects of substance and form that are particularly useful for this small

[1] A full overview of genre key issues and definitions can be found in Boudouride's excellent and thorough literature review (Boudourides, 2001)

feasibility study of the Wikipedia domain. First of all, however, it would be helpful to look at Ecologies, which perfectly describe how these texts evolve and are modified in this domain.

2.2 Ecologies

Duff (2000, 4) pointed out that due to the existence of biological metaphors in genre theory (Erickson, 2000; Kwasnic and Crowston, 2005), it was only natural that the evolutionary paradigms found in Darwin's "Origin of Species" (Darwin, 1859) would be used to model the ways in which literary forms change over time by evolving, being modified and being replaced. Duff (2000, xii) also suggested that some genre theorists also extend the biological metaphor in "quasi-Darwinian terms" by describing some of the mechanisms of literary evolution as "the competition of genre", genres struggling for survival, their "fitness" for an environment and the "possibility of extinction" but this could be criticised for extending the metaphor too far. Kwasnic and Crowston (2005, 87) gave an impressive description of how the genres behave when they extended Erickson's genre "ecology" metaphor (Erickson, 2000). They compared a genre to an organism in an ecological community: they all rely on other organisms for their effectiveness, have an effect on each other, evolve over an unspecified course of time at different paces, and can even replace each other, i.e. memo-genre. They declared that these ecological habitats are CoPs (see Section 2.2), Wikipedia, in the context of this paper. As is the case in most areas of research, however, there are issues with Web genres that have to be considered when studying digital media such as Wikipedia. Kwasnic and Crowston (2005, 87) described these issues and how the problems arise in a genre ecology by explaining two phenomena which occur more or less concurrently: firstly, traditional genres appearing on the Web and, secondly, the appearance of new unique genres appearing on the WWW. Both of these phenomena have genres that divide, merge, transform and evolve. This is an important implementation issue that has to be taken into consideration because the genres have to be identifiable by all systems and perceptible to all users.

2.3 Communities of Practice

Wenger (2000) stated that CoPs are social institutions or sites where human agents draw on genre rules to engage in organizational communication which operate by producing, reproducing, or modifying whatever they are producing (in this case, genres). (Yates and Orlikowski, 1992, 305) stated: "In structurational terms, genres are social institutions that are producing, reproducing, or modifying when human agents draw on genre rules to engage in organizational communication". If the behaviour of the community could be comprehended, this could be exploited in the implementation of skimming and categorization tools that would provide search and retrieval of important community objects. Further to this, Collins et al. (2001) explained that what the community sees as important will be reflected in the implicit structures found in the objects they create and share and as Watt (2009) has observed "convergence on a set of standardised

document structures is both natural and helpful". These objects are genres that occur in the web; CoPs are utilised, but we need to look at the ways in which these web pages are structured in Wikipedia and the types of features of which they consist. Wenger (2000) described what he considered to be the characteristics of a CoP as:

The community: In pursuing their interest in their domain, members engage in joint activities and discussions, help each other, and share information. They build relationships that enable them to learn from each other. A website in itself is not a CoP. Having the same job or the same title does not make for a CoP unless members interact and learn together.

The Domain: A CoP is not merely a club of friends or a network of connections between people. It has an identity defined by a shared domain of interest. Membership therefore implies a commitment to the domain, and therefore a shared competence that distinguishes members from other people.

The Practice: They develop a shared repertoire of resources: experiences, stories, tools, ways of addressing recurring problems-in short a shared practice. This takes time and sustained interaction.

3 Evolution of Wikipedia Genres

Wikipedia is an important and popular domain for accessing information about a huge range of information. Not only do individuals use it for reference, but many organisations, such as the BBC News, use it for information. However, Wikipedia does have its detractors, who criticise it for being inaccurate; it suffers from vandalism, of course, which is carried out sometimes with malicious intent, but also sometimes just to raise a laugh. The Now Show (British comedy program) on BBC Radio 4 has even used Wikipedia for some of its sketch material.At a higher level there are many types of offshoots of Wikipedia such as WikiBooks (Cookbooks, StudyGuide etc,), Wikizine, Portals etc. However, this study concentrates on the evolving types in 'Wikipedia The Free Encyclopaedia'. This Wikipedia operates in an editorial hierarchy of: all, users, Autoconfirmed users, Bots, Administrators, Bureaucrats, Checkusers, Stewards and Board Vote Administrators with least permissible editing powers being assigned to "all" and "users" and the most 'power' to "Stewards" and "Board Vote Administrators". For example, once an edit is submitted 'live' by a least empowered editor, a modification is accepted/rejected by Stewards et al. The full hierarchy and list of responsibilities is published in Wikipedia but will not be listed here. Wikipedia contributors are allowed to edit each page and are given a toolbox of HTML functions to use for text formatting, linking files, adding photographs, inserting tables and so on. Much like Kwasnic and Crowston (2005) describe traditional genres are appearing on the web. The Wikipedia community, we believe, contains a wide array of such types, such as FAQ, lists for example list of films etc, Reviews, Guides, News Articles, Events and so on. Not only that, new unique genres also appear, transform and evolve, much like

Kwasnic and Crowston (2005) pointed out. Section's 3.1-3.3 will be used to examine how some of these structural forms (or genres) evolve. It could be argued that Wikipedia (encyclopaedic) itself could be called a genre in its own right but for this study, we look at the articles (maybe sub-genres?) of which the content and form are constantly evolving as a result of editors employing certain devices or tools, such as formatting text, lists, tables and photographs and also studying multiple sources, such as biographical books, for amending and adding factual content. Underlying each article in Wikipedia there is also a discussion area (or aka Talk Pages) between users that re-enforces our potential understanding of the whole CoP aspect of this domain. For example, much of the current discussion about General Motors Corporation (see Figure 3) is the likelihood of its demise in the current financial crisis and debate about what content to include. The Wikipedia site says the purpose of the talk pages is to provide areas for editors to discuss changes to the linked article or project page. Also provided is a history from when the article was first created to the present day as each amendment no matter how big or small is recorded. This small study is overall being used to examine the suitability of Wikipedia for our study into structural forms and how structure is perceived and used by purpose and form. Our overall goals, at this stage, are to examine the suitability of Wikipedia and its constituent parts (discussion etc) as a vehicle for demonstrating the CoP and evolutionary paradigm in this context in which we have devised a methodology (3.1 below). We have chosen to look at the evolution of several possible new and old types of structured articles (see Figures 2, 3, 4 and 5) such as discographies, lists musical groups/bands, footballers etc as well as conduct a small case study of how a Wikipedia biographical article such as Spike Milligan evolves.

3.1 Methodology

The methodology for this study consists of several parts which tie in with the Ecology, CoP and the Orlikowski and Yates (1994, 544-545) definition of purpose and form.

1. Search: REM, Margaret Thatcher, General Motors and Alex Ferguson of Manchester United Football Club etc
2. Examine the potential genres by purpose and form.
3. Look at how the articles are constructed and note if they lead to any other types of structure (Kwasnic and Crowston, 2005) such as discography, FAQ, Biography, List and so on. Look at the articles, noting in particular whether:
    a) They are traditional types of genre such as Biography.
    b) The article is a NEW style of genre.
    c) Examine the underlying CoP to see whether the discussions (in articles) indicate the expected characteristics indicated (Wenger, 2000).

3.2 Search and Record Genres

The Wikipedia articles were first perceived for their potential usefulness during the relevance judgements ('paper' exercise) for the INEX in 2006 (Huang et al., 2006). While examining the topical relevance of organisations' submissions during the relevance judgements' phase of INEX 2006, it was noticed that particular structures or genres were starting to appear throughout. This showed that Wikipedia would be a potentially suitable vehicle for studying the evolution or development of genre in a CoP and also for studying highly visual types of text with perceivable purpose and form. After the search by subject most of the important types of articles linked to the main articles were mapped, recorded and analysed. As all the genres could not be mapped out due to space issues, they have been narrowed down to internal categories such as: biographies, lists, football clubs, motor vehicle manufacturers, and political parties which have their own particular purpose and form. Conducting the search enabled the recording of the relevant statistics, purpose and form attributes that are shown in Table 2.

A popular area in modern culture is, of course, music such as rock and pop. While searching for rock music, it was noticed that there was a hierarchy of genres which are connected to musical groups such as REM, Muse, etc (Figure 2) which link to other types of genres such as discography, biography, musical group, several types of lists list of bands under the same record label, chronological list of Rock and Roll Hall of Fame inductees which is in two forms. One list [2] has a large table with the band information containing the year order, name, image of artist and year inducted and the second list type is in alphabetical order (Table 2 has more information).

Figure 2 shows that there are some already existing web genres in Wikipedia such as list and index but also new ways of structuring information. The Musical Group, Band member and Discography contains a layout consisting of lists and tables but some titles also show up consistently throughout different examples of Musical Groups (U2, Muse, etc), Discographies and Band Members. It is also clear that, similarly to the evolutionary paradigms in section 2.1, some of these literary forms are evolving, being modified and being replaced. Some of the existing genres are actually evolving and outliving their usefulness and in some circumstances leading to a new type, for example, the histories of the articles for rock bands REM and U2. Three years after the original articles had appeared, they seemed to have become too big and thus seemed to have outlived their usefulness. The editors created other articles, such as discographies[3], to help contain the information, leaving the textual information laid out helpfully for the readers who were then able to filter to the content they would most need. This is particularly helpful in an information search task. As can be seen in Figure 3 and Table 2 Automobile Manufacturers, such as General Motors and Ford, had several different types of articles linked to the main result.

At the top of the hierarchy, the Automobile Manufacturers could be categorised as an Organisation (for example, British Petroleum and GM Corporation nearly have similar

[2] http://en.wikipedia.org/wiki/List_of_Rock_and_Roll_Hall_of_Fame_inductees

[3] http://en.wikipedia.org/w/index.php?title=Talk:R.E.M._discography\&oldid=94780788

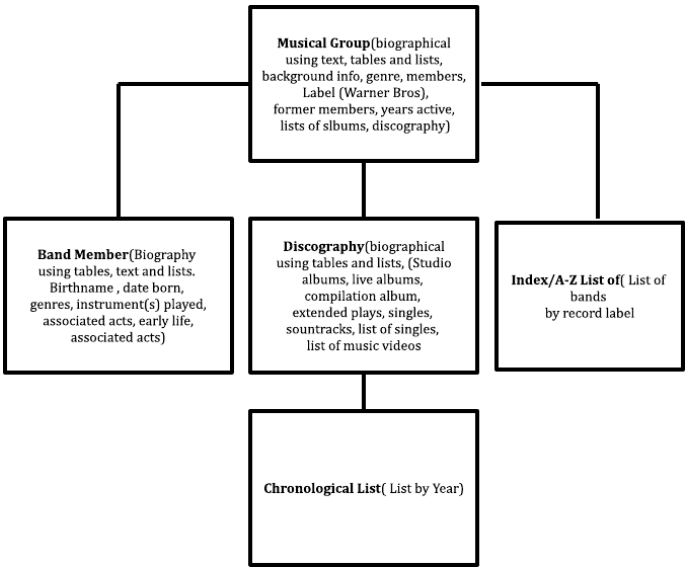


Figure 2: Band/Music Group example (see REM for a good example): visual format from Wikipedia.

structures) and display a particular structural form that allows the perceiver/reader to understand and find quickly the appropriate content pertaining to the organisation or Automobile Manufacturer. By emphasising the most important information related to each article, for example, in figure 3 (also Table 2) the community of editors has decided that the most important information defining an Organisation (such as General Motors) are what you see in the boxes above (as well as an image of the Organisation logo). This information is heavily formatted due to its prominence in the article whereas the rest of the article is composed only of text and citations of a biographical nature that elaborate on this information. Wikipedia has many articles on particular organisations in the automobile industry, such as GM Corporation, British Leyland and Ferrari (Figure 3). In the next level of the hierarchy, the first two organisations are more famed for producing consumer or family cars whilst the latter, Ferrari, produces Formula 1 or SuperCars (Figure 3). The SuperCar and Family Car have their own individual forms, but occasionally overlapping, attributes, such as, an engine. During the analysis of the biography genre, it was noticed that several types of biography exist along with links to their genres.

There was another type of biographical sub-genre or, arguably, mixed genre found: Football Manager. This structured article also naturally led to Football-Player, Team and Ground, which also linked to County and Country. The football team/club article seemed to outgrow its purpose and lead to new genres such as manager, ground and

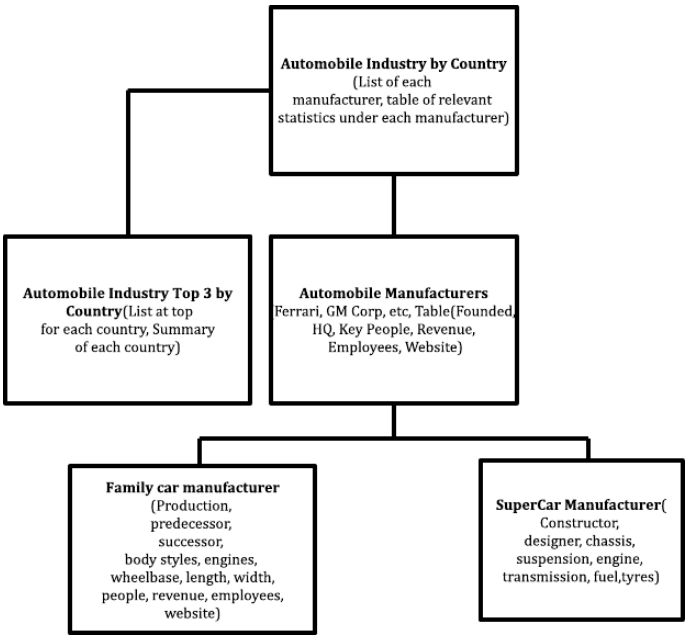


Figure 3: Small search for Automobile related Wikipedia articles and analysis of how they are structured and linked.

player. As Figure 4 (table 2 for more information) shows, each genre type is defined by certain forms that have been created in this particular community.

Search for football related Wikipedia articles and analysis of how they are structured and linked (see also Table 2). Several different types of biographical genre exist in Wikipedia with many different sets of characteristics for some notable figures in history, such as, Spike Milligan, Nelson Mandela, Alex Ferguson (Manchester United manager), Pol Pot, John Howard, Karl Marx and Margaret Thatcher. Other than the sole biographical structures for Spike Milligan, Karl Marx et al., a different form existed for ex-prime Minister John Howard, ex-president Nelson Mandela and ex-prime Minister Margaret Thatcher which, as can be seen in Figure 5, contains particular layout titles along with a biographical 'substance' in chronological order – this genre could be called: Leader. Many kinds of genres that are represented by several types of structure and meaning have been recorded in figures 2-5. Table 2 lists most of these recorded types and shows the attributes according to which we would contend they qualify to be categorised by form and purpose. An examination of the related interactions on the discussions pages and edits of the articles mentioned above showed that Wikipedia can qualify as a CoP because it contains the three characteristics outlined by Wenger (2000):

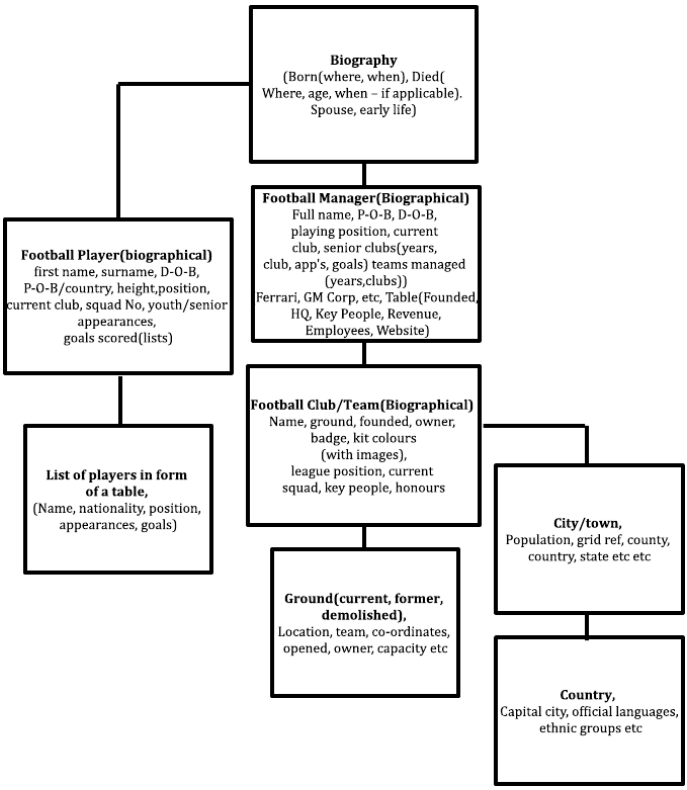


Figure 4: Search for football related Wikipedia articles and analysis of how they are structured and linked (see also Table 2).

The Practice, The Community and The Domain. The editors involved demonstrate a commitment to the domain and also seem to value their collective competence and the chance to learn from each other. The members engage in joint activities, such as voting, interaction and discussion. The editors develop a large and shared repertoire of resources, such as stories, tools and ways of addressing recurring problems, a mechanism for this being that the editors actually practice democracy by initiating voting cycles to discuss the merits of carrying out an alteration to an article[4].

[4]Vote Proposal: `http://en.wikipedia.org/w/index.php?title=Talk:R.E.M.\_discography\&oldid=94780788`

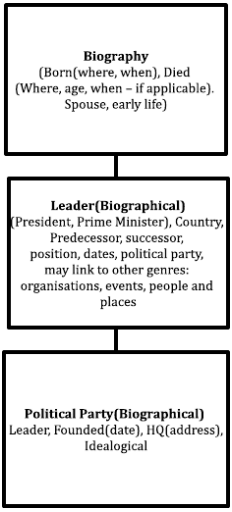


Figure 5: Search for notable people in history and analysis of how the main and related articles are structured and linked.

3.3 A Closer Look at Genres

An analysis of the literature on ecologies, CoPs and Yates and Orlikowski (2002, 15) work will help us to further identify the ways in which articles are created in domains such as Wikipedia. Also, by referring to the history of articles being made available, we can find out the details of how and when the particular articles (or genres!) are/were produced, reproduced and modified. Although it could be argued that carrying out an analysis of the edit histories, discussion/talk pages would, instead of demonstrating genre evolution, simply suggest the supplementing of previous knowledge or thoroughness, this would be a narrow-minded view of the genre evolution. The analysis of the edit histories and discussion clearly indicate a CoP implementing the division, merging, transformation and evolution of the article genres in this very complex domain. We looked closely at an example of a 'classical' genre, the biographical article, in this case about Spike Milligan, the celebrated and highly influential comedian and author who died in 2002. The purpose of the article is, obviously, to provide biographical information to the reader about Spike Milligan. As can be seen, the form of this web page article is continuously evolving and being transformed, much as Kwasnic and Crowston (2005) described in their ecological metaphor. The original article was first created on November 5, 2001; note the sparse and poorly organised information it contains (Figure 6).

After seven years, approximately 487 different users have submitted edits to the Spike Milligan page with only ten editors submitting more than 10 edits per person. The community for this particular article is evidently quite large and, as can be seen in

Terence Alan "Spike" Milligan (1918-) 'Irish' comedian, novelist, poet, and member of the Goons. Spike Milligan has suffered from Bipolar Disorder for most of his life.

Comedy shows:

* The Goon Show
* Q8

Books:

* Puckoon
* Adolf Hitler, My Part in his Downfall

Resources

* `http://www.fireflycafe.org/spike/`
* `http://www.google.com/Top/Arts/People/M/Milligan,_Spike/`

Figure 6: Spike Milligan Wikipedia article containing no formatting or notable structure dated 5 November 2001.

Table 1, the article has grown considerably. Over this time period, images were placed within the article. Eventually, the portrait picture in Figure 7 (after being in many different positions) ended up at the top right as nearly all pictures now do. On the 26th November 2006 a table with the title *Spike Milligan* was created by a contributor.

Since the screenshot was captured in early 2008, the biographical form in Figure 7 has yet again been transformed after much discussion by the editors involved. Not only has the contents table on the left been extended, but the table that encapsulates the name has also changed. The Birth name, Born, Died and Children information has now changed to Born, Died, Nationality, Influences and those people on whom he arguably exerted an influence. The focus is now concentrated on the career instead of on the person, much as with the Football Player or Leader, and is thus maybe moving towards forming another unique kind of genre which we could rename Artist or maybe Comedian. There could also possibly be overlaps with one of the new genres with classical forms, such as Obituary (as Milligan has died) and Biography. Another possible issue which could be linked with the merging and overlapping of genres is the reaching of a consensus on what constitutes a type of genre in a community, in this case a biography. Recently, the Spike Milligan article has evolved to contain more professional information than biographical (a human life in its course). The main elements in Figure 8 (below), "Children," has been amended to show professional influences and those whom he influenced instead of children, spouses (some time ago). The available 'histories' and underlying discussion area (Talk Page) do, of course suggest this but the information is not conclusive. It is obvious that by operating as a community, the contributors have added and enhanced information that they deem important (in a

Figure 7: Table 'feature' located in top-right of each biographical article.

hierarchy of importance) and have placed extra structural emphasis on the elements which are deemed most important about each article genre even if they do not always agree on these details. We noticed, by examining the history and discussion areas, that the Wikipedia editors have utilised a toolbox of HTML functions for formatting and embedding various media links, such as, video and photographs. The editors also seem to access unlikely sources to obtain information as indicated by one editor in the Talk page discussion: an un-named 'source' in the Daily Telegraph is cited as possessing a photograph of Spike Milligan's gravestone (for inclusion in the article) which is famous for the Gaelic inscription: "I told you I was ill".

4 Conclusion and Future Work

This paper constitutes the first steps in research on the Wikipedia domain, in particular, how the structures evolve in this "organic" and "biological" type of community. Wikipedia seems to be a suitably large and hierarchically structured CoP to demonstrate how genres evolve over a scale of time which will allow us to look closely at the evolution of a biography even though not all articles featured in Wikipedia are as formed as others. The viewed articles also contain a good fusion of form and purpose although some of the less formed articles contain a very small amount of form. The next step in this research is to formulate a study on genre and perception in this new area, that is, Wikipedia, which has the same aims and objectives as described in the earlier research paper by Clark (2007), and in the electronic mail study of Clark et al. (2008). A particular user search study will be set up to complement further research by looking into how the Wikipedia articles are used and perceived when a user extracts the form and recognises the purpose of the documents during an information search. The plan is to study how human beings cognitively interact and use genres of documents, which features or attributes they perceive and whether their perceptive processes can be explained or understood. Users are typically asked to read and categorise material from different genres and with different structures and forms. Measuring user categorisation according to genre, structure and form is further enhanced by recording eye movements during the tasks. Detailed data can thus be obtained regarding the attention paid to

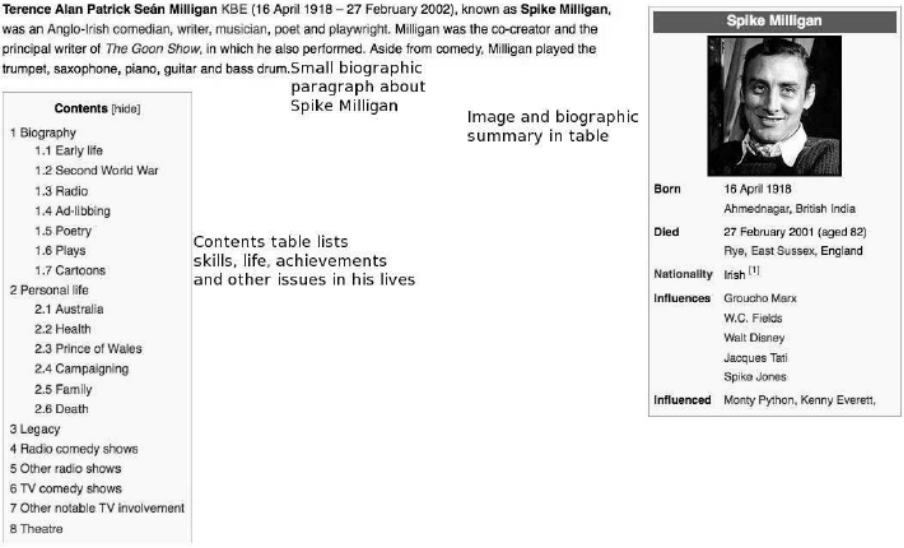


Figure 8: Biography example: visual format from Wikipedia containing tables, lists and image dated early 2008.

structures and forms by users when recognising, judging and determining genre. This research has the potential to show how human categorisation behaviour can be emulated computationally by a machine that actually 'understands' the meaning of a text for automatic retrieval. In some contexts, in particular, it is important to find out which of the two predominant processes – ecological (perceiving for action and affordances – cf. Gibson 1986) and constructivist (perceiving for recognition – cf. Gregory 1966) – are present in the subjects' genre recognition tasks.

References

Aristotle (1984). *The Rhetoric and the Poetics of Aristotle*. Modern Library College Editions Series. McGraw-Hill Higher Education, 1st edition.

Berkenkotter, C. and Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: cognition, culture, power*. L. Erlbaum Associates, Hillsdale.

Boudourides, M. A. (2001). Commorg topics of genre literature review. Unpublished Article.

Breure, L. (2001). Development of the genre concept. *(last checked = 2009-05-22)*. `http://people.cs.uu.nl/leen/GenreDev/GenreDevelopment.htm`.

Clark, M., Ruthven, I., and Holt, P. O. (2008). Genre analysis of structured emails for corpus profiling. In *Proceedings of the Workshop on Corpus Profiling for Information Retrieval and Natural Language Processing*. EWICS.

Clark, M. J. (2005). Classifying xml documents by genre vol. 1. Master's thesis, The School of Computing.

Clark, M. J. (2007). Structured text retrieval by means of affordances and genre. In *BCS IRSG Symposium: Future Directions in Information Access BCS IRSG Symposium: Future Directions in Information Access BCS IRSG Symposium: Future Directions in Information Access BCS IRSG Symposium: Future Directions in Information Access*. British Computer Society.

Collins, T. D., Mulholland, P., and Watt, S. N. K. (2001). Using genre to support active participation in learning communities. In *In: The European Conference on Computer Supported Collaborative Learning (Euro CSCL 2001)*, Maastricht.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Dewdney, N., VanEss-Dykema, C., and MacMillan, R. (2001). The form is the substance: classification of genres in text. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Duff, D. (2000). *Modern genre theory*. Longman Publishing Group, London, 1st edition.

Ehmann, K., Large, A., and Beheshti, J. (2008). Collaboration in context: Comparing article evolution among subject disciplines in wikipedia. *First Monday: Peer-Reviewed Journal on the Internet*, 13(10).

Emigh, W. and Herring, S. C. (2005). Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, volume 4, page 99a. IEEE.

Erickson, T. (2000). Making sense of computer-mediated communication: Conversations as genres, cmc systems as genre ecologies. In *Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 3*. IEEE Computer Society.

Gibson, J. J. (1986). *The ecological approach to visual perception*. LEA, New Jersey, 2nd edition.

Giddens, A. (1986). *The Constitution of Society: Outline of Theory of Structuration*. University of California Press.

Gregory, R. L. (1966). *Eye and Brain the psychology of seeing*. World University Library, London, 1st edition.

Halliday, M. A. K. (1973). *Explorations in the Functions of Language*. Edward Arnold (Publishers) Ltd.

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London.

Huang, F., Watt, S., Harper, D. J., and Clark, M. (2006). Robert gordon university at inex 2006: Adhoc track. In *Overview of INEX 2006*, volume 4518/2007, pages 64–72. Springer-Verlag.

Ingwersen, P. and Jarvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Information Retrieval Series*. Springer, 1st edition.

Kress, G. and Threadgold, T. (1988). Towards a social theory of genre. *Southern Review*, 21(3):215–243.

Kwasnic, B. and Crowston, K. (2005). Introduction to the special issue genres of digital documents. *Information Technology and People*, 18(2):76–87.

Lalmas, M., Rolleke, T., Szlavik, Z., and Tombros, T. (2004). Accessing xml documents: the inex initiative. In Agosti, M. and Fuhr, N., editors, *DELOS WP7 Workshop on the Evaluation of Digital Libraries*, University of Padua, Italy.

Martin, J. R. (1999). Mentoring semogenesis: 'genre-based' literacy pedagogy. In Christie, F., editor, *Pedagogy and the Shaping of Consciousness: Linguistic and social processes*, Open Linguistics Series, pages 123–155. Cassell, London.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, 70(2):151–67.

Orlikowski, W. J. and Yates, J. A. (1994). Genre repertoire: norms and forms for work interaction. *Administrative Science Quarterly*, 39:541–574.

Watt, S. (2009). *Text categorisation and genre in information retrieval.* (In Press), chapter In Press. John Wiley and Sons.

Wenger, E. (2000). Communities of practice and social learning systems. *Organization*, 7(2):225–246.

Yates, J., Orlikowski, W. J., and Rennecker, J. (1997). Collaborative genres for collaboration: Genre systems in digital media. In *Proceedings of the 30th Hawaii International Conference on System Sciences: Digital Documents - Volume 6*, volume 6, pages 50–59. IEEE.

Yates, J. A. and Orlikowski, W. (2002). Genre systems: Structuring interaction through communicative norms. *Journal of Business Communication*, 39(1):13–35.

Yates, J. A. and Orlikowski, W. J. (1992). Genres of organizational communication: a structurational approach to studying communication and media. *Academy of Management Review*, 17(2):299–326.

Table 1: Article Structure by Contents Table (positioned top-left of each article see Figure 4 Left Hand Side) Evolving bi-annually

Nov 2001	Nov 2004	Nov 2006	Nov 2009
Comedy Shows: ∗The Goon Show ∗Q8 Books: ∗Adolf Hitler, My Part in his Downfall ∗Puckoon	∗1 Biography ∗2 Radio Comedy Shows ∗3 TV Comedy ∗4 Theatre ∗5 Movies ∗6 Books ∗7 Quotations ∗8 External Links	∗1 Biography ∗2 Posthumously ∗3 Trivia ∗4 Radio Comedy ∗5 Other radio shows ∗6 TV comedy shows ∗7 Theatre ∗8 Films ∗9 Books	∗1 Biography o 1.1 Early Life o 1.2 WW II o 1.3 Radio o 1.4 Ad-libbing o 1.5 Poetry o 1.6 Plays o 1.7 Cartoons ∗2 Personal life o 2.1 Australia o 2.2 Health o 2.3 Prince of Wales o 2.4 Campaigning o 2.5 Family o 2.6 Death ∗3 Legacy ∗4 Radio comedy shows ∗5 Other radio shows ∗6 TV Comedy Shows ∗7 Other TV ∗8 Theatre ∗9 Films ∗10 Books ∗11 Quotations ∗12 External links ∗13 References

Table 2: Article, Genre, Purpose and Form

Genre	Stats (Date Created/Amount of Editors/Edits)	Attributes of Purpose (Themes, topics, discourse structure)	Attributes of Form(Structural features e.g titles, lists etc)

Band/Musical Group (Query REM)	1 February 2002, 1067 editors ,1564 edits	To biographically present the past and present members of the group, show their work output and list their achievements.	∗TABLE TITLES, HEADINGS: Background information, Origin Genre(s) ,Years active Label(s), Associated acts, Website(URL), Former members. MAIN TEXT HEADINGS Chronological History, URL(s)to listen/download radioone or more song samples, Summary of the Discography.
War (query Napoleonic Wars)	22 March 02, 991 editors, 2361 edits		TABLE TITLES/HEADINGS date, location, result. List of Beligerants, names of sides, List of commanders on each side, casualities and losses on each side in numerics. MAIN TEXT HEADINGS (title and years of stage) Lead up to start of war, major phases of war(battles etc), outcome, legacy and effects.
Discography(query REM)	17 December 2005/ 145 editors /410 edits	To present and list the output produced by an entity such as musical artists TABLE TITLES, HEADINGS Small summary table with type of release and amount e.g. album 5.	MAIN TEXT HEADINGS: Tables. Each table by title such as Studio Albums, Singles etc with sub-titles such as Year, Album and Single Details, chart positions.
A to Z index List of Bands (by genre query alternative rock bands)	27 March 2004, 832 editors, 4520 edits	To present a comprehensive alphabetically structured index of alternative musical groups throughout the world.	TABLE TITLES, HEADINGS Contents table 0 to 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z MAIN TEXT HEADINGS Small summary. Index of alphabetical sections with list of bands name beginning with 0 to 9 to Z.
Leader query Nelson Mandela	7 June 2005, 317 editors, 1053 edits	To present the biographic details of how and when a person became a leader in a political party etc	TABLE TITLES HEADINGS Title of office held, dates held position, Vice president, succeeded and/or proceeded by, born where and when, political party MAIN TEXT HEADINGS Early life, key moments in life and leadership

City query Aberdeen	5 February 2002, 28 editors, 2234 edits	To detail the geographic, population and historical information pertaining to a particular city.	TABLE TITLES HEADINGS Name of city, map with location, population, density, language spoken, location Council area, Lieutenancy area, Constituent country, Sovereign state, Post town, Postcode district Dialling code, Police or fire ambulance(name of service, European Parliament, UK Parliament Scottish Parliament MAIN TEXT HEADINGS Geography, demography, climate, Landmarks, transport, culture
Football Club query Scarborough Athletic	25 June 2007,68 editors,451 edits	To present current and historical information, including achievements, regarding a football and/or soccer team.	TABLE TITLES HEADINGS image of coat of arms, Full name, Nicknames, Founded, Ground (Capacity), Owner, Managing Director, League, Premier League. Images of club strip(shorts, socks and top). MAIN TEXT HEADINGS: Stadiums, Supporters, Table of honours, records, Table with list of current aquad players. Tables (with lists by name and years) coaching staff, key people, manager history, chairman history.
List of Football Players query List of Newcastle United F.C. players	11 February 2006, 110 editors,357 edits	To present current(still playing) and historical information(now retired), including achievements, regarding a football or soccer player.	TABLE TITLES/ HEADINGS None MAIN TEXT HEADINGS Large table with headings Name Nationality, Position, Club Name career, appearances, Goal Table with list of first team captains (year and name)
List of Lists query list of bands by genre etc	10 December 2003/195 editors/340 edits	A comprehensive list of lists sorted by certain categories.	TABLE TITLES None. MAIN TEXT Title(By Genre, By Instrument) then list under each
Political Party query socialist party of Ireland	12 February 2004/92 editors/553 edits	Presents biographic information regarding a political party in a particular country or region in the world.	TABLE TITLES/ HEADINGS Name, Logo, Founded, Leader, Headquarters, Political ideology, International Affiliation, European Affiliation European Parliament Group, Colours , Website. MAIN TEXT HEADINGS: Electoral history, Key policies, List of elected members(name, position, district)

Automobile Manufacturer (query General Motors)	25 February 2002/1772 editors/5233 edits	Presents information to the public regarding the general business structure and financial performance.	TABLE TITLES/HEADINGS Type , Founded, Founder(s) Headquarters, Area served, Key people, Industry , Products, Services , Revenue ▽ currency (year), Operating income ▽ currency(year), Net income ▽ currency(year), Total assets ▽ currency (year), Total equity ▽ currency (year), Employees, (number)(year), Divisions, Subsidiaries, Website (url) MAIN TEXT HEADINGS: History, Company Overview, Corporate Structure, Table listing open manufacturing plants, Table of Yearly Sales, List of brands/defunct brands, subsidiaries

Philip M. McCarthy, John C. Myers, Stephen W. Briner,
Arthur C. Graesser, Danielle S. McNamara

A Psychological and Computational Study of Sub-Sentential Genre Recognition

Abstract

Genre recognition is a critical facet of text comprehension and text classification. In three experiments, we assessed the minimum number of words in a sentence needed for genre recognition to occur, the distribution of genres across text, and the relationship between reading ability and genre recognition. We also propose and demonstrate a computational model for genre recognition. Using corpora of narrative, history, and science sentences, we found that readers could recognize the genre of over 80% of the sentences and that recognition generally occurred within the first three words of sentences; in fact, 51% of the sentences could be correctly identified by the first word alone. We also report findings that many texts are heterogeneous in terms of genre. That is, around 20% of text appears to include sentences from other genres. In addition, our computational models fit closely the judgments of human result. This study offers a novel approach to genre identification at the sub-sentential level and has important implications for fields as diverse as reading comprehension and computational text classification.

Key words: Genre recognition, reading comprehension, text classification.

Introduction

The term *genre* designates a category of text (Graesser, Olde, & Klettke 2002). As with all categories, a genre cannot be specified by a qualitative analysis of a single exemplar (Davies & Elder 2004), but rather reflects the characteristics of a family of exemplars. A genre has an underlying set of norms that are mutually understood (either consciously or unconsciously) by the audience for whom the text was created (Downs 1998; Hymes 1972). Thus, it is the presence, prevalence, and prominence of the norms characterizing the genre of a text that allow the text to be recognized as an *interview*, a *lecture*, a *conversation*, a *story*, a *home page*, a *blog*, an *exposition* of some aspect of science, history, or art, or any other genre from a myriad of possibilities.

Any definition of genre would assume that the text in question is of a sufficient length for it to be classified on the bases of the features that accrue. Interestingly, we know of no study below the paragraph level in which genre has been deemed recognizable. Further, because genre has traditionally been viewed as a characteristic of the text (Biber 1988, Graesser et al. 2002), there is the implicit assumption that the texts in a genre have some degree of homogeneity. The common features of genre may be either absolute invariance (that is, the feature is necessary for a genre), but more frequently they are statistical reg-

ularities (i.e., the feature occurs more frequently in one genre than alternative genres). Whether the features are absolute or statistically distinctive, however, there is the question of how much and what type of information is needed to make a classification decision that a text is in a genre. That is, if a text T belongs to a genre G then the text itself is composed of sub-textual features (i.e. phrases, clauses, sentences) that are always or frequently diagnostic of genre G.

In this study, we investigate these assumptions by collecting data that explore six primary questions:

(1) How short (in terms of number of words) can a text be for its genre to be accurately recognized?
(2) What types of errors (i.e., genre misclassifications) do readers make when identifying genres?
(3) To what degree are texts heterogeneous (i.e., have characteristics of multiple genres)?
(4) Does the process of genre identification depend on reading skill?
(5) What textual features (e.g., syntax, lexical choice) influence genre identification?
(6) Can a computational model categorize genre using only as much text as humans appear to need?

Psychological and Computational Goals of Study

This study serves two primary purposes: one psychological, relating to reading comprehension; and one computational, relating to text classification.

Reading comprehension. Readers' comprehension of a text can be facilitated or otherwise influenced by the text genre, which is identified on the basis of the textual characteristics (Bhatia 1997; Graesser et al. 2002; Zwaan 1993). Given that familiarity with textual structure is an important facet of reading skill, training struggling readers to recognize text structure can help students improve their comprehension (Meyer & Wijekumar 2007; Oakhill & Cain 2007; Williams 2007).

Available research in discourse processing indicates that skilled readers utilize different comprehension strategies that are sensitive to text genre (van Dijk & Kintsch 1983; Zwaan 1993). Once a text genre is identified, it guides the reader's memory activations, expectations, inferences, depth of comprehension, evaluation of truth and relevance, pragmatic ground-rules, and other psychological mechanisms. For example, when reading a history text, it is important to scrutinize whether an event actually occurred. In contrast, in most narrative fiction, the truth of the event is not a particularly relevant consideration (Gerrig 1993), presumably because there is a "willing suspension of disbelief" (Coleridge 1985). Further, expository texts are more likely to discuss unfamiliar topics. Consequently, the lack of sufficient prior knowledge forces higher ability

readers to process the details of the text at a more local level (e.g., connections between adjacent clauses). In contrast, narratives are more easily mapped onto everyday experience and, as a result, readers tend to process the global and thematic relationships in a passage (Otero, Leon, & Graesser 2002). Empirical evidence supports such claims through recall (Graesser, Hoffman, & Clark 1980) and reading time experiments (Graesser, Hoffman, & Clark 1980), demonstrating that narrative text is recalled approximately twice as well as expository text, and also read approximately twice as fast. Thus, stylistic surface structure attributes of the language and discourse vary in importance dependent upon the genre of the text (Zwaan 1993).

A better understanding of the nature of text genre is important for text comprehension theories as well as interventions to improve comprehension. If readers are using different strategies to process different genres of text, then it is important to understand the processes and information constraints during the course of genre identification. An understanding of the circumstances under which readers make correct or incorrect attributions of genre could expand our knowledge of the reading strategies used for each genre.

Text classification. According to the Netcraft Web Survey (December, 2007), the internet consists of *at least* 155,230,051 sites, an increase of 5.4 million sites since the previous month. And with many sites boasting 1000s of web pages, the number of web documents available to browsers is astronomical. With such an abundance of information, locating the desired information is becoming ever more problematic.

Search engines categorize web pages using *spiders*, which crawl through the internet, storing information embedded in web pages. Although each spider is different, the typical information gathered from web pages is based on high frequency words, key words in headers and links, and meta-tags that specifically indicate terms of relevance. But despite such a broad approach (or maybe *because of* such a broad approach), the majority of web pages located from any given search are not relevant to the user. This overabundance of non-relevant documents is generally caused by such search criteria establishing not the genre of the text (e.g., blogs, home pages, narratives) but the *topic* of the text (e.g., politics; see Boese, 2005, McCarthy, Briner, Rus, & McNamara, 2007; Santini, 2006)

One approach to narrowing user searches is to locate documents based on genres (Stamatatos, Fakotakis, & Kokkinakis, 2001), and particularly web genres (Meyer zu Eissen & Stein, 2004; Roussinov et al., 2001; Santini, 2006). Definitions of *web* genres do not differ substantially from definitions of *text* genres. For instance, Boese (2005) argues that web genres are elements of the presentation of the article, effective analyses of the writing style, the formats or layouts of the documents, and the actual content of the articles. Roussinov et al. (2001) argues that web genres have socially recognized norms of format and purpose that appear in the text. Whereas traditional text genres may include

expository, interview, conversation, and children's story, web genres subsume text genres and include others such as home page, opinion, review, course description, and blog.

The advantage of genre categorization over (or in addition to) topic categorization is one of focus. For instance, a Google search for the leading candidates in the 2008 presidential race for the White House (as of March, 2008), returned 1000s of web pages on the relevant *topic* (e.g., current affairs in the presidential race), but included *genres* as diverse as news, blog, review, research group, TV archive, and Q&A site. While all such genres may potentially provide the user with the desired information, it is safe to assume that most searches would be facilitated by the option or the availability of classification by genre.

A better understanding of the nature of web genres is important for search classification approaches. Improved knowledge of what constitutes a genre can lead to improvements in the efficiency of spiders. As a result of such improvements, categorizing searches by genres can help users by limiting and focusing the returned web pages, offering significant savings in time and effort.

Our approach: Less is more?

Within our six research questions, our *approach* to genre recognition focuses on the following two questions, previously unaddressed in the genre literature: (1) How long does a text have to be for it to be considered a member of a genre? And (2) To what degree are text genres heterogeneous; that is, is a text of one genre composed entirely of sentences that are also identified as being of that genre?

Regarding the first question, we can safely assume that a person who reads an entire book, article, or web page will have little doubt as to its genre. Similarly, we can assume that the first word alone from such a reading might not inspire great confidence that the correct genre will be identified. Our question is how much text is necessary for most readers to accurately identify the genre of a text.

Our first question is important in any model of genre identification, and comprehension in general. The sooner the reader identifies the genre of the text, the sooner the appropriate background knowledge will be activated and guide comprehension accordingly. We can also hypothesize that readers who recognize a text's genre earlier and more accurately possess more developed reading skills. That is, their experience or knowledge better allows them to recognize genre specific words or structures. Thus, it is conceivable that early and accurate genre recognition may be a diagnostic, practical estimation of reading skill.

Our second question regards the heterogeneity of text. While a given text T may be considered a member of genre G, we cannot assume that G is wholly composed of sentences from G. For instance, a science text may begin with a scene setting *narrative*, or a *history* of the theme to be considered. Similarly, a

web blog may comprise (and indeed must comprise) *news* as much as *views*. Such a conjecture is highly related to our first question; that is, if sentences (or sub-sentences) are recognizable as genres, then what is the distribution of such genre-recognizable-fragments across text? A better understanding of the composition of genres may facilitate improving reading comprehension. For example, if lower grade-level science texts contained more narrative sentences (presumably a form more familiar to the readers) then the expository information in the text may be more easily integrated.

Just as both our main questions address reading comprehension, they also address computational text classification. While categorizing web searches may facilitate users by focusing web page returns, the additional processing of documents may be prohibitive. Franklin (2008) reports that Google search engines operating at peak performance, using four spiders, could crawl at a rate of 100 pages per second. While such a performance is impressive, *peak* performance is not *typical* performance and with many billions of pages to crawl through, many billions of seconds are required. However, processing time can be significantly reduced by limiting the amount of text needed to be analyzed. For instance, early approaches to genre classification assessed the text as a whole (Biber 1988; Karlgren & Cutting 1994; Kessler, Numberg, & Shütze 1997), and this tradition continues into contemporary studies (Boese 2005; Bravslavski & Tselischev 2005; Finn & Kushmerick 2006; Kennedy & Shepherd 2005; Lee & Myaeng 2002, 2004; Meyer zu Eissen & Stein 2004; cf. Lim et al. 2005, for results on titles and meta-tags). But by considering genre as an identifiable feature at the sub-sentential level, perhaps only a small amount of text needs to be processed. If so, identifying the genre of a text and the heterogeneity of the text may be feasible with a relatively small (possibly random) sample.

The Experiments

This study includes three experiments. Experiments 1 and 2 constitute the psychological portion of the study, focusing on our first four research questions: (1) How short (in terms of number of words) can a text be for its genre to be accurately recognized?; (2) What types of errors (i.e., genre misclassifications) do readers make when identifying genres?; (3) To what degree are texts heterogeneous?; and (4) Does the process of genre identification depend on reading skill? Experiment 3 constitutes the computational portion of the study, focusing on our remaining two research questions: (5) What textual features (e.g., syntax, lexical choice) influence genre identification?; and (6) Can a computational model categorize genre using only as much text as humans appear to need?

Experiment 1

The goal of Experiment 1 is to investigate experts' ability to recognize the genre of sentence fragments presented out of context. Specifically, we examine wheth-

er three experts in discourse psychology agree on the genre classification for isolated sentence beginnings; and, if so, how many words are required for accurate genre classification to occur. Experiment 1 is limited in the number of participants because what might be described as our first *real question* is simply "is the task even possible?" Given that numerous psychological and computational studies have investigated genre using text no shorter than the paragraph, it is appropriate that our initial study is relatively modest in scope.

The Genres

In this experiment (and throughout the study), we consider three genres: *narrative*, *history*, and *science*. We include science and narrative because they have been the focus of numerous previous psychological studies (e.g., Albrecht, O'Brien, Kendeou & van den Broek 2005; Linderholm & van den Broek 2002; Mason, & Myers 1995; Kaup & Zwaan 2003; Trabassao & Batolone 2003) and, therefore, provide a relatively uncontroversial point of departure. We include history because whereas no one disputes that science texts can be described as expository, there is a question as to whether history is more expository-like or more narrative-like. Some researchers, for example, have recognized that history texts can be similar to narratives, the two genres tending to be presented more as a chronological series of events on topics with which many readers are familiar (Duran, McCarthy, Graesser, & McNamara, 2007; Tonjes, Ray, & Zintz 1999). In contrast, other researchers (e.g., Radvansky, Zwaan, Curiel, & Copeland 2001) have used history texts as examples of expository texts, without any mention that such a genre could be considered narrative-like.

Empirical computational approaches to distinguishing the genres used in this study provide evidence for both categorizations: For instance, McCarthy, Graesser, and McNamara (2006) used an array of cohesion indices showing that history texts were more similar in structure to science texts. That is, both history and science texts were more cohesive than narrative texts. On the other hand, Duran et al. (2007) used temporal indices and found evidence that history texts were more similar to narratives. That is, both history and narrative texts were structured similarly in terms of temporal development. Meanwhile, Lightman, McCarthy, Dufty, and McNamara (2007) found evidence for all three genres having distinct characteristics. Thus, one question addressed in this study is whether history sentences are correctly classified to a similar degree as narrative and science sentences; and if not, to which genre are they more likely to be assigned. As such, the choice of genres used throughout this study was motivated by two considerations. First, that the genres were sufficiently diverse in terms of structure, style, and purpose that differences in recognition accuracy would be identified; but second, that distinguishing the genres would not be a trivial task.

Predictions

For the narrative genre, we predicted that incorrectly assessed sentences would more likely be classified as history sentences because both genres typically describe past events. For the history genre, we predicted misclassified sentences to be equally distributed between narrative and science, because history texts are equally likely to be descriptive of an event (thus, narrative-like) or feature explicit lexical cause and effect relationships (thus, science-like). For the science genre, we predicted that misclassified sentences would more likely be assessed as history sentences, because some elements of scientific texts present explanations from a chronological perspective.

We further predicted that our expert raters would correctly identify a high percentage of sentences requiring approximately only half of the words in a sentence to do so. This prediction is based on typical features of verb and pronoun positioning. Verbs, for example, feature early in a sentence, and their tense is indicative of their genre (McCarthy et al., 2008). Similarly, the subjects of sentences are generally positioned at the beginning of sentences. Regardless of whether the subject of the sentence is a pronoun or named entity, the characteristics of the sentence subject are at least somewhat indicative of text genre.

Corpus

The corpus in our analysis was composed of a subset of sentences taken from the 150 academic text corpus compiled by Duran et al. (2007). In that corpus, the texts were sampled from 27 published textbooks provided by the MetaMetrics repository of electronic duplicates. A subset of the Duran and colleague's corpus (McCarthy et al., in press) further focused the corpus by filtering out an equal number of similarly sized paragraphs. The McCarthy and colleague's sub-corpus featured 207 paragraphs in total (828 sentences): 69 paragraphs in each of the three genres, and 23 paragraphs each of 3, 4, and 5 sentences in length. The approach we adopted for sentence selection from these paragraphs is based on studies indicating that topic sentences are processed differently to other sentences in a paragraph (e.g., Kieras 1978, Clements 1979, McCarthy et al. in press). Because such research also indicates that topic sentences are more likely to occur in the paragraph initial position (Kieras 1978; McCarthy et al. in press), we sampled an equal number of paragraph-initial sentences and paragraph-non-initial sentences. For the paragraph-non-initial sentences, we used the third sentence of each paragraph. This choice was made for two reasons. First, all paragraphs contained a third sentence; and second, third-sentences are presumably less closely related in terms of co-reference to first-sentences than first-sentences are to second-sentences; thus, the effects of a possible confound are reduced. This reduction to first-sentences and third-sentences left 414 candidate sentences in our corpus. To ensure that participants viewed sentences of approximately equal length, we further reduced the size of the corpus by only including all sentences that were within one SD of the average length in terms of number of words of the 414

candidate sentences (mean number of words = 15.437; SD = 7.113). Using this criterion, 298 sentences remained, of which the smallest group was 35 sentences belonging to the genre of narrative-paragraph-non-initial. We thus selected 35 to be the number of sentences from each of the six groups (narrative/history/science by paragraph-initial/paragraph-non-initial). Consequently, our corpus consisted of 210 sentences, equally representing the three genres and the initial/non-initial sentence dichotomy (see Appendix).

Method

Participants. The participants included three researchers in discourse processing (one post-doc, one graduate student, and one advanced and published under-graduate). Each participant assessed each of the 210 sentences that equally represented the genres of narrative, history, and science.

Procedure. A Visual Basic program was created to evaluate genre recognition. The program included three parts: *instructions*, *practice examples*, and *testing*. Following the instructions, participants were provided with six practice sentences. Once the practice was completed, a message informed the participants that the experiment would begin. Each participant evaluated all 210 sentences. The sentence order was randomized for each participant. The program operated by displaying the first word of the first sentence in a text window. Participants were required to assess the genre to which they thought the sentence fragment belonged. Participants registered their choice by clicking on one of four on-screen buttons: *Narrative*, *History*, *Science*, and *Don't Know*. As soon as a genre choice was made, the next word from the sentence appeared in the text window. All punctuation was retained in the display and was attached to the word it adjoined (e.g., in the sentence fragment *Yes, it was a …* the word *Yes* would appear as *Yes* + comma).

After 10 seconds, if the participant made no decision, then a new word automatically appeared in the text window with a message informing the participant of the new word. The variables of *genre choice* and *accuracy* were recorded. Participants evaluated each word of each sentence until they had either given the same decision of the genre of the sentence three consecutive times (whether right or wrong), or until all the words in the sentence were presented. The final choice of participants was recorded as the genre choice, regardless of previous decisions. For the variable *number of words*, the number was determined as the point of the first instance of a choice in a string of three consecutive identical choices. Thus, if a participant's genre selection was *don't know*, *don't know*, *narrative*, *science*, *science*, *science* then the count at the point of the first instance of *science* would be the number of words used: in this case, four words. That is, although the participant viewed six words in total, the partici-

pant's final choice occurred at the fourth word and was confirmed by the fifth and sixth selections.

Results

Raters

We begin our analyses by demonstrating inter-rater reliability. This reliability establishes confidence in our evaluation of the data as typical of expert ratings and is particularly important when using few raters. On average, the raters correctly identified the genre of the sentences for 90% of the data. Inter-rater agreement between Raters 1 and 2 for correctly assessed sentences was approximately 90% ($X^2 = 41.077$, $p < .001$). Inter-rater agreement between Raters 1 and 3 was also approximately 90% ($X^2 = 47.569$, $p < .001$). And the Inter-rater agreement between Raters 2 and 3 was approximately 91% ($X^2 = 61.145$, $p < .001$).

Of the 210 sentences assessed, *all three* raters classified the correct genre for approximately 69% of data. Two of the three raters correctly classified an additional 17% of the sentences (i.e., 86% of the data). At least one of the three raters correctly identified an additional 6% of the data (i.e., 92% of the data). Also, less than 9% of the data were incorrectly assessed by any of the raters. Thus, the raters' accuracy was quite high. Further reliability of the raters' analyses can be demonstrated in terms of recall and precision (see Table 1). Such accuracy and agreement between the three raters (M=82%) offers support for the forthcoming analyses to be considered representative of genre recognition at the word level by experts in discourse processing.

	Accuracy			Correct			Misclassification			
	Recall	Precision	F1	Narrative	History	Science	Narrative	History	Science	DK
Rater 1	.824	.840	.832	.914	.829	.729	.081	.052	.023	.019
Rater 2	.824	.892	.856	.871	.857	.743	.062	.038	.000	.076
Rater 3	.810	.817	.813	.886	.757	.786	.081	.076	.024	.029
Mean	.819	.850	.834	.890	.814	.752	.075	.055	.016	.041

Table 1: Accuracy and misclassifications for Narrative, History, and Science texts, and "Don't Know"(DK) classifications.

Genre

In the experiments presented throughout this study, the accuracy of the results is reported in terms of *recall*, *precision*, and *F1*. Such reporting is common when, as in this study, we are concerned with predictions of categories (i.e., narrative, history, science). To briefly explain each term, *recall* (R) shows the number of

correct predictions divided by the number of true items in the group. In other words, recall is the number of *hits* over the number of *hits + misses*. *Precision* (P) is the number of correct predictions divided by the number of correct and incorrect predictions. In other words, precision is the number of *hits* divided by the number of *hits + false alarms*. The distinction is important because an algorithm that predicts everything to be a member of a single group will account for all members of that particular group (scoring 100% in terms of recall) but will also falsely claim many members of other group(s), thereby scoring poorly in terms of precision. Reporting both values allows for a better understanding of the accuracy of the model. The F1 value is the harmonic mean of precision and recall. It is calculated as *2PR / (P+R)*.

In terms of genre recognition accuracy, the expert raters correctly classified 516 of the 630 sentences (i.e., 210 sentences * 3 raters): an average accuracy of 82% (see Table 2). This result is in line with our prediction. While the results appear consistent across the genres (Min. F1 = 82, Max. F1 = 84), closer analyses suggest that the genres elicit quite distinct patterns of responses.

	Decisions		Accuracy			Misclassifications			
Domain	Selected	Correct	Recall	Precision	F1	Narrative	History	Science	DK
Narrative	234	187	0.890	0.799	0.842	/	10	3	10
History	206	171	0.814	0.830	0.822	25	/	7	7
Science	168	158	0.752	0.940	0.836	22	25	/	5

Table 2: Accuracy and misclassifications of expert raters by domain for Narrative, History, and Science texts, and unclassified "Don't Know" (DK) texts

Narratives. The narrative genre received the highest recall value (89%); however the narrative genre was also the least precise (80%), with 47 additional false alarms. Indeed, of all misclassifications, more sentences were incorrectly assigned by the experts as narrative, than either of the two expository genres (narrative = 51%; history = 38%; science = 11%). The misclassifications to the narrative genre suggest that narrative sentence structures may be the most ubiquitous type. The approximately equal division of false alarm narrative sentences to the science (22) and history (25) genres further suggests that the two expository genres may comprise, to a small but notable degree, narrative-like sentences. Indeed, for six sentences (three history and three science) all three-raters categorized the sentences as narratives (see Table 3).

Example	Domain	Sentence
1	History	We cannot[1] sell the lives of men and[3] animals[2], said one Blackfoot chief in the 1800s, „therefore we cannot sell this land."
2	History	I[1] had vainly[3] flattered[2] myself that without very much bloodshed it might be done.

3	History	Much to[1] my surprise[2], I[3] had forgotten my glasses in prison, so I used my wife's.
4	Science	Taking no joy[1] in life, looking forward[3] to nothing, wanting to withdraw from people and activities[2].
5	Science	This, he thought[1], would[2] demonstrate[3] that emotions can be mechanically induced (Cohen, 1979).
6	Science	Watson[1,] 3 went even[2] further and suggested that at the human level, deep emotions are also just the result of association and learning.

Note: The superscript number indicates the point at which the genre selection was made

Table 3: The six sentences identified by all raters as narratives.

Looking more closely at these "misclassified" sentences, we observe that all three raters classified Example 1 as narrative by the 9th word of the sentence. It is only after this point that the words *Blackfoot chief* reveals the sentence more clearly as a history text. For Example 2, all three raters classified the text by the 4th word. Indeed, although the text recounts an historical event, the use of first person pronoun (rare in expository structures) may be indicative of a narrative style of writing. This appears again in Example 3. All three raters classify the sentence in Example 3 by the 5th word. Again, the incorporation of first-person pronouns renders the sentence more narrative-like, even though the text as a whole is taken from a history book. Example 4 is actually a sentence fragment and resulted in one rater having to view the entire sentence before deciding that it was narrative[1]. While the sentence lists symptoms of depression, the text could easily be read as describing a character. For Example 4, all raters agreed on narrative by the 5th word. However, had the raters read a little further, the science-like nature of the sentence (passive construction) may have been more easily recognized. The final example is deemed narrative by the 3rd word. It is possible that the raters saw the subject word *Watson* and considered the text to be from Sherlock Holmes. The results are in line with our predictions that the early presence of key lexical and grammatical features triggers the expert readers' genre recognition.

History As predicted, when history sentences were misclassified, they tended to be identified as narratives. This result supports the conclusions of Duran et al. (2007) and Tonjes et al. (1999). The three examples above (see Table 3) demonstrate the type of narrative-like text that appears to be a feature of history texts.

Science Only 75% of the science sentences were classified accurately, the lowest of the three genres. However, when raters did label a sentence as from the science genre then they were nearly always correct to do so (precision = 94%, the highest of the three genres). Of the 52 misclassified science items, most were

1 This sentence was subsequently modified for later experiments.

attributed to history (25) and narrative (22). The high history value is as predicted, because much scientific discussion begins from a historical perspective. The equally high narrative value suggests that science texts may be equally viewed as narrative-like in the description of many of their topics.

Don't Know As predicted, the raters correctly identified the vast majority of items. Only 22 sentences remained unclassified with no particular domain attracting more *Don't Know* classifications. Only one sentence was rated as *Don't Know* by all three raters: *Many of those years were harsh and cruel.* Although from a history text, the sentence could equally well be attributed to narrative given that the author seems to be voicing an opinion rather than an objective fact.

Number of Words Used

High inter-rater reliability is required to establish confidence that the number of words used by raters to assess the genre of sentences is suitably representative of experts' judgments. Following Hatch and Lazarton (1991), the adjusted correlation for three raters was $r = .660$, $p < .001$. For items for which *all three* raters correctly assessed the genre of the sentence, the correlation was $r = .732$, $p < .001$. The consistency across raters means that we can take the average number of words used by raters as the gold-standard representative of experts in assessments of the genre of sentences.

For the corpus as a whole ($N = 210$), the average number of words used by raters was 4.948 ($SD = 2.818$; Mode = 5). As predicted, this is less than half the average length of sentences in the corpus; indeed, it was a *third* of the length. However, when we divide the corpus for the condition of *all raters giving correct judgments/other sentences*, the results show that significantly fewer words were required to *correctly* identify the genre (Correct: $N = 144$, $M = 4.419$, $SD = 2.407$; Incorrect: $N = 66$, $M = 6.101$, $SD = 3.256$; $F(1,208) = 31.140$, $p < .001$, $\eta^2 = .130$). This result suggests that a rater judgment of *fewer* than five words is more likely to be correct, and a judgment of *greater* than five words is more likely to be *in*correct. The three sentences for which raters took the most words to arrive at the *wrong* genre are shown in Table 4.

Domain	Classification	Sentence
Narrative	Don't Know	Friends in the barrio explained that the director was called a principal, and that it was a lady and not a man.
History	Narrative	The governor presided over an advisory council, usually appointed by the governor, and a local assembly elected by landowning white males.
History	Don't Know	We blow the whistle that's heard round the world, and all peoples stop to heed and welcome it.

Table 4: The three longest, misclassified sentences.

To better understand the above result, we considered each genre individually. The results suggested that the five-word average applied only to narratives (Correct: $N = 187$, $M = 4.808$, $SD = 3.029$; Incorrect: $N = 23$, $M = 7.870$, $SD = 4.808$; F (1, 208) = 18.028, $p < .001$). There was no significant difference for correctly identifying genre using fewer words for the genres of history or science. The similarity here between the history and science genres and the distinction from narrative genre offers support to the conclusions of Graesser et al. (2002), McCarthy et al. (2008) and McDaniel et al. (1986). The result offers evidence that if an expert reader of a narrative sentence has not become sufficiently aware of the sentence's genre by the fifth word that it is unlikely that subsequent words will make the reader any the more sure of the genre.

Discussion

In Experiment 1, we asked three experts in discourse processing to identify the genre of isolated sentences culled from a corpus of narrative, history, and science texts. Demonstrating high agreement, the raters showed that expert readers could significantly identify the genre of over 80% of sentences. Further, our raters demonstrated that fewer than five words (less than a third of the sentence) were required to correctly classify these sentences. Indeed, for the narrative sentences, viewing more than five words did not improve the accuracy of identifying the genre. These results suggest that the first third of sentences alone contains sufficient genre characteristics for skilled readers to begin the process of activating knowledge of text structure: a process which facilitates comprehension.

Our results also showed that expert readers viewed many of the history and science sentences as narrative, suggesting that expository texts tend to comprise a notable number of narrative-like sentences. On the other hand, regardless of the genre from which sentences were taken, our raters were least likely to classify sentences as science. This result sheds like on the heterogeneous compositionality of text, providing significant implications for computational research in genre recognition. Specifically, computational approaches to genre recognition have tended to assume that the text as a whole is representative of the genre or text-type to which it has been assigned (e.g., Biber 1988, Louwerse, McCarthy, McNamara, & Graesser 2004). The results of Experiment 1 suggest that texts of any given genre may typically comprise sentences from many other genres. Understanding this diverse compositionality may lead to changes in how computational tools assess text searches and evaluations.

The compositionality of text is also a factor for research in reading development. Our results here suggest that for a text to be suitably representative of any given genre, it may require that the text contains a notable number of sentences more indicative of other genres. If a text does not contain this mixture of genre sentences, it is possible that a reader may have greater difficulty processing the text, as certain expectations may not be met.

In Experiment 1, we also addressed the question as to whether the genre of history was closer to science or to narrative. Our results suggest that expert readers are as able to identify and distinguish history sentences as they are science and narrative sentences. This result supports the findings of Lightman et al. (2007), who found that history texts were distinct from both science and narrative texts. However, if we consider only the 39 misclassified sentences of the history genre, our results showed that 64% of these sentences were incorrectly assigned by our experts as narratives, whereas only 18% of the sentences were identified as science (and the remainder as *don't know*). Viewed this way, the result suggests that a notable portion of history texts comprise narrative-like structures, a result that supports Duran et al. (2007), who found that history texts were more narrative-like than science-like. The categorization of history texts is important to cognitive science as many experiments have assumed that a history text is an expository text (e.g. Radvansky et al. 2001). Consequently, researchers can often assume that history text will lead to *similar* results as science text and *different* results from narrative texts. The results of Experiment 1 demonstrate that such an assumption could lead to erroneous conclusions.

Above all, the results of Experiment 1 demonstrate that genre recognition at the sub-sentential level is possible. There having been no previous investigations of how much text is required to recognize genre, this first experiment indicates that very little text is actually required and that readers most likely activate information about text structure very early in the reading process. Such recognition might provide a signature of reading ability, and as a consequence, a method of assessing reading ability. The principle results of Experiment 1 certainly provide sufficient initial evidence that such an approach is viable and that this paradigm can be further explored as an assessment of reading skill. In addition, if only the first five words of a sentence is sufficient for experts to recognize the text's genre, then computational approaches to text analyses may need to follow this lead. That is, text assessment for such features as readability, difficulty, cohesion, and genre recognition may also need to be performed on just the first third of sentences because it is here that a significant portion of human evaluation of the text seems to occur. More specifically, computationally evaluating an entire sentence may incorrectly assess the sentences' remaining two-thirds as relevant to the reader's processing. Indeed, this remainder may be redundant or even noise in terms of reader activation of certain processing components. In Experiments 2 and 3 we explore these issues more closely.

Experiment 2

In Experiment 1, three experts (i.e. published authors) in discourse processing were asked to identify the genre of isolated sentences culled from a corpus of narrative, history, and science texts. The experts had high inter-rater agreement (min = 90%) and required about a third of the words in the sentence to accurately identify genres (accuracy as measured by F1, Narrative = .82; History =

.84; Science = .82). The results further showed that these experts often classified history and science sentences as narrative, suggesting that expository texts tend to be composed of a notable number of narrative-like sentences. On the other hand, science-like sentences were the least likely to be misclassified into other genres, suggesting the science-like sentences seldom occur in the non-science genres. The results also showed that these skilled readers required about a third of the sentence to successfully activate sufficient knowledge to recognize textual genres. Presumably, this activation skill is beneficial to reading and comprehension development. As such, we might expect that the number of words necessary to correctly recognize genres to be indicative of reading ability.

The results of Experiment 1 were intriguing. However, the most compelling result was the one informing us that genre recognition at the sub-sentential level was, indeed, possible. To establish greater confidence in our paradigm, Experiment 2 builds on Experiment 1 by including a larger sample of participants, an independent assessment of reading ability, a measure of *time on task*, and recording accuracy in terms of *number of words* used. In this experiment, we ask four main questions. First, how quickly (in terms of number of words) do readers identify the genre of a text? Second, what types of errors (i.e., genre misclassifications) do readers make when identifying genres? Third, does the process of genre identification depend on reading skill? And fourth, how does *time on task* affect the accuracy of genre decisions?

Corpus

The corpus used in Experiment 2 was the same as that used Experiment 1, with the following modification: We modified one science sentence that was a sentence fragment, changing *Taking no joy in life, looking forward to nothing, wanting to withdraw from people and activities* to *Examples are taking no joy in life, looking forward to nothing, wanting to withdraw from people and activities*.

Methods

Participants. There were 22 participants (Male = 10, Female = 12; *M* = 24.1 years old) who received $50 in exchange for participation in two experiments, of which, this was one. The other experiment was unrelated. All participants were native English speakers. Fifteen participants were undergraduate students, five participants were graduate students, and two participants identified themselves as non-students.

Assessments. To assess reading skill, we used the Gates-MacGinitie (GM) reading test, a multiple-choice test consisting of 48 questions designed to measure reading comprehension. We used the level 10/12 version of the test, which has a reliability of .93 (MacGinitie et al, 2002).

Participants' genre recognition was evaluated using a similar Visual Basic program to that used in Experiment 1. Three variables were recorded: *genre*

choice, *accuracy*, and *time on task*. To accommodate the *time on task* assessment, the following modification from Experiment 1 was made: As in Experiment 1, participants made their selection by clicking on one of four on-screen buttons: *Narrative*, *History*, *Science*, and *Don't Know*. However, in Experiment 2, the buttons' position was randomized such that the genre choice could appear in any of the four buttons. Upon selecting one of the buttons, the mouse cursor returned to a central position so that each button was always equidistant from the start point of the cursor. As soon as a genre choice had been made, as in Experiment 1, the next word from the sentence appeared in the text window.

Results

Subject Analysis

Our results showed that participants typically needed only a sentence's first three words to make their decision on genre (overall words used: $M = 3.35$, $SD = 1.50$; words used in correct assessments only: $M = 3.33$, $SD = 1.45$).The average accuracy of genre categorization was high (Recall: 0.86; Precision: 0.71; F1: 0.77), and this accuracy was consistent across the three genres (see Table 5). These results are consistent with Experiment 1.

Genre	Accuracy	Mean	SD
Narrative	Recall	0.86	0.09
	Precision	0.71	0.12
	F1	0.77	0.09
History	Recall	0.71	0.14
	Precision	0.76	0.09
	F1	0.72	0.11
Science	Recall	0.67	0.12
	Precision	0.88	0.09
	F1	0.75	0.11

Table 5: Accuracy of genre evaluation

While the average number of words used for correct assessments was 3.33, the mode for number of words used in correct assessments was 1.00 (25.02% of the data, see Table 6). The second highest frequency for number of words used was 2.00 (21.88%), followed by 3.00 (15.36%), and so forth such that the distribution of words used for correct assessments described a logarithmic curve ($df = 16$, $F = 244.95$, $p < .001$, $r^2 = .939$). Such a result is unlikely to mean that participants blindly hit the same genre choice button, because the genre buttons randomly changed position, meaning that participants had to find their genre

choice. Additionally, the result is unlikely to suggest that participants were simply trying to get the task done as quickly as possible because examining *all* final decisions made on the first word (in other words, decisions for which participants had selected a genre on the first word and selected that same genre for the second and third words), 50.69% of the genre decisions were correct (baseline = 33.34%). As such, there is some evidence here that humans make their genre decision on the very first word of a sentence, and more often than not their decision is correct.

The magnitude of the correlation between *reading skill* (GM) and *words used* was moderate ($r = .37$, $p = .09$), as was the relationship between *words used* and *accuracy* (in terms of correlations with F1 participant evaluations, Science: $r = .43$, $p < .05$; Narrative: $r = .37$, $p = < .09$, History: $r = .37$, $p < .09$). We examined the results more closely by dividing the participants into two groups based on a mean split of the Gates-MacGinitie test scores ($M = 24.00$; $SD = 9.14$). Using these values, 13 participants were designated as lower-skill (LS) and 9 participants were designated as higher-skill (HS). Differences in Gates-MacGinitie test scores were analyzed using Levene's test for equality of error variances. No significant differences between groups were detected ($p > 0.5$), indicating that the groups are suitable for comparison.

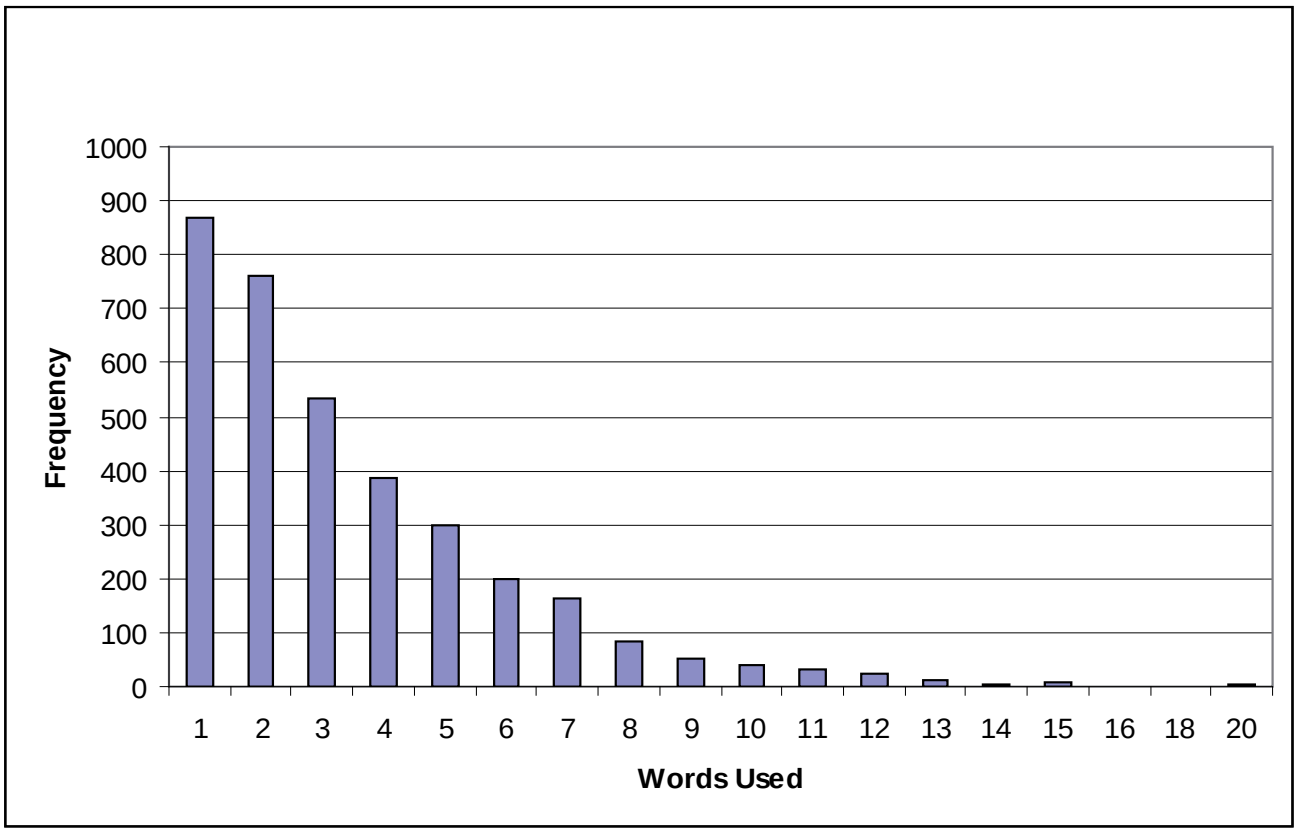


Table 6: Frequencies of number of words used in correct genre assessments.

We conducted an exploratory Analysis of Variance (ANOVA) to determine which of 22 variables best distinguished the reading skill groups. The analysis revealed that 7 variables significantly distinguished the two skill groups ($p < .05$) and 4 variables were marginally significant ($p < .10$; see Table 7).

	Low skill		High Skill				
Dependent Variable	Mean	SD	Mean	SD	F	P	η^2
Narrative precision	0.66	0.12	0.79	0.08	7.55	0.01	0.27
Time: 3rd word (History)	1.01	0.29	0.72	0.21	6.72	0.02	0.25
Science Recall	0.62	0.13	0.74	0.06	6.52	0.02	0.25
Science F1	0.71	0.11	0.81	0.07	5.87	0.02	0.23
Time: 3rd word (Narrative)	0.96	0.29	0.70	0.16	5.75	0.03	0.22

Table 7: Five most predictive variables in distinguishing low/high skill readers

The *narrative-precision* variable suggests that higher-skilled readers tend to be better at *not* classifying non-narrative sentences as narratives. In other words, skilled readers know better when a sentence is *not* a Narrative. These readers' greater accuracy may be because they are prepared to use more words than the lower-skilled readers. However, a t-test revealed no significant differences between the number of words required by lower-skilled readers ($M = 2.97$; $SD = 1.21$) and higher-skilled readers ($M = 3.85$; $SD = 1.68$), $t > 1.0$, $p > .1$. Despite the lack of a significant difference between the higher-skilled and lower-skilled readers in terms of words used, the direction of the difference suggests that lower-skilled readers may too easily assume the direction or nature of the sentence discourse.

The variable, *time on task for the 3rd word in history sentences,* indicates the time on task for judging the third word of history sentences for correct decisions. Lower-skilled readers took significantly *more* time on this word. Indeed, *time on task* negatively correlated consistently with GM reading skill across all three genres for both 2nd words of sentences (Narrative: $r = -.427$, $p = .05$; History: $r = -.443$, $p = .04$; Science: $r = -.523$, $p = .01$) and 3rd words of sentences (Narrative: $r = -.596$, $p < .01$; History: $r = -.606$, $p < .01$; Science: $r = -.500$, $p = .02$). These results suggest that higher-skilled readers may be able to more quickly integrate new information.

Taken together, the results suggest that higher-skilled readers are more able to quickly and accurately process sentential information, using as few as the first three words. This advantage appears most evident in two features: on the 3rd word of sentences (all other word positions demonstrated weaker results); and in the precision result for the narrative genre. One further variable of interest is that higher-skilled readers may be prepared to use more words before making genre decisions. This final point is consistent with Experiment 1 in which expert readers (and therefore, presumably higher in ability than those who participated in

this Experiment) tended to use at least two more words than those who participated here. However, caution should be taken with this conclusion because a step-wise multiple regression revealed that only the time on task for 3rd words of history sentences variable contributed to the model (adjusted R-square = .336).

Item Analysis

Of the 210 sentences in Experiment 2, only 4 (2%) failed to be correctly evaluated by any of the participants. For instance, the history sentence "*I had vainly flattered myself that without very much bloodshed it might be done"* was evaluated by all participants as a narrative; and the science sentence "*Hindi is the most widely used, but English is often spoken in government and business"* was evaluated by 20 participants as history and by 2 as narrative. A further 33 sentences (16%) were correctly categorized by all the participants. For instance, the narrative sentence "*Why, I wouldn't have a child of mine, an impressionable little thing, live in such a room for worlds*" resulted in no misclassifications. For over half the sentences (55%) at least 19 of the 22 participants correctly evaluated the genre. For instance, the science sentence "*In areas with hard water, many consumers use appliances called water softeners to remove the metal ions"* recorded only three misclassifications. Conversely, only 10% of the sentences received less than 6 correct evaluations, an example being the narrative "*The Empress of Russia looked dressed for war, Igor thought.*"

The item analysis also showed that the sentences that received the highest accuracy in terms of categorization were likely to require fewer *words* for such categorization to be made. Thus, there was a negative correlation between the percentage of participants who correctly evaluated a sentence and the number of words needed to correctly categorize the sentence (r = -.639, p < .001). For example, "*Chemical weathering processes change the chemical composition of rocks"* was correctly identified as a science sentence by all of the participants and required an average of only 1.23 words to be identified. In contrast, "*However, this process was too slow to satisfy the Renaissance demand for knowledge and books"* was correctly categorized by only 27% (n = 6) of the participants and required 10 words to be correctly identified as a history sentence.

The results of the *time on task* demonstrated similar results. Specifically, there was a negative correlation between the percentage of participants who correctly assessed a sentence and average *time on task* for assessment (r = -320, p < .001). The results for both *words used* and *time on task* were consistent across the genres of narrative (words: r = -.613, p < .001; time: r = -.466, p < .001); history (words: r = -.701, p < .001; time: r = -.404, p < .001); and science (words: r = -.578, p < .001; time: r = -.257, p = .034).

Thus, consistent with the results of Experiment 1, viewing more words does not lead to greater genre classification accuracy. This result indicates that if a sentence does not contain genre-specific features early in its structure, then it is also unlikely to contain those features later in its structure. The results for *time*

on task indicate that sentences that are more accurately classified are also more quickly classified. We can presume that the quicker the decision, the less the processing necessary to make the correct decision. Thus, we did not observe a time/accuracy tradeoff.

Collectively, the results suggest that most sentences from the three genres can be accurately categorized in relatively few words and relatively little time. However, the variation within this accuracy suggests a continuum of *sentence-categorization difficulty*. That is, the first few words of sentences can often be sufficiently non-prototypical or ambiguous to reduce the likelihood of correct reader categorization. As such, it is feasible that the construction of the initial aspects of a sentence may significantly affect sentence processing, with less prototypical constructions causing readers to activate less relevant expectations of prior knowledge.

Discussion

In Experiment 2, 22 participants identified the sentence genres of 210 sentences. The results indicated that both higher- and lower-skilled readers used about three words to accurately identify genres. Two primary variables related strongly to participants' reading ability: *Narrative-precision* and *Time on Task* for the 3rd word (i.e., typically the word with which participants make their decision). Thus, higher-skilled readers are less likely to think a sentence is a narrative when it is not, and they also require less time to make their decisions.

Taken together, the results of Experiments 1 and 2 allow us to make the following conclusions. The results suggest that 1) a wide range of readers can accurately categorize genres at the sub-sentential level; 2) as few as the first three words of a sentence may be all that is required for that assessment to occur, and in over half the cases just the very first word; 3) genre recognition may be indicative of reader ability; and 4) variables such as *time on task*, *accuracy*, and *number of words used* may be the indicators of reading ability.

The research presented in these initial two experiments offers an interesting and promising direction toward a better understanding of how genre knowledge is represented in the mind and subsequently activated. We plan to use this understanding to better establish our *genre identification paradigm* as an assessment of reading skill, and even as a possible intervention for reading development. While much remains to be done in this respect, the results presented here offer an exciting new perspective on the nature of text and the possibilities of reading skill assessment.

Experiment 3

Introduction

The results of Experiments 1 and 2 provided evidence that genre recognition could be accomplished with a high degree of accuracy using as few as the first

three words of sentences. Given such accuracy from such little discourse information, we can hypothesize that readers are utilizing shallow lexical and syntactic sentential features to identify genre. To address this hypothesis, we examined whether a computational model based on only lexical and syntactic features (i.e., the information apparently used by participants) provided similar results. If the model could replicate the results found with humans, then it potentially provides evidence that participants use such sentential features when processing text.

In Experiment 3, we construct a computational model based on our results from Experiments 1 and 2. We use the model to investigate what information could be present in the initial words of sentences such that it can provide participants with sufficient information to make a genre evaluation. The question of whether or not we could build a computational model is important for two reasons. First, our computational model sheds light on the features of the text that most likely influences readers' genre classifications. And second, if a computational model can categorize genre using minimal sentence information, then such an approach could facilitate text classification systems.

Computational Approaches to Text Classification

Computational approaches to categorizing genre have tended to treat text as a homogeneous whole. Thus, the whole text is analyzed and, based on the results, the text is categorized as a single genre. Such an approach is as common in traditional text-genre classifications studies (e.g., Biber 1987; Biber 1988; Duran et al. 2007; Hall, McCarthy, Lewis, Lee, & McNamara 2007; Karlgren & Cutting 1994; Louwerse et al. 2004; McCarthy, Graesser, & McNamara 2006; McCarthy, Lewis, Dufy, & McNamara 2007) as it is in web-genre classification studies (e.g., Boese 2005; Bravslavski & Tselischev 2005; Finn & Kushmerick 2006; Kennedy & Shepherd 2005; Lee & Myaeng 2002, 2004; Meyer zu Eissen & Stein 2004).

For example, in traditional genre classification studies, Biber (1987) identified lexical diversity and singular person pronoun use as key predictors in distinguishing British-English from American-English. Kessler, Nunberg, and Schutze (1997) used part of speech tags, lexical cues (e.g., Mr. and Mrs.), punctuation features, and shallow discourse features such as sentence length, to distinguish registers such as editorials, romantic fiction, and biographies. Louwerse et al. (2004) used cohesion values to distinguish both spoken from written texts and narratives from non-narratives. And Stamatatos, Fakotatos, and Kokkinakis (2001) used various style markers such as punctuation features and verb- and noun-phrases frequencies to distinguish between the authors of a variety of newspaper columns. What each of these studies have in common is that the whole text is analyzed and, based on a distribution of features, is labeled as a member of a single category.

Meanwhile, the more contemporary web-genre identification studies typically rely on three categories of features: *style*, *form*, and *content* (Boese & Howe,

2005). Style includes readability formula (e.g. Flesch Kincaid Grade Level), syntactical information (e.g. passives/actives), and various heads of phrases such as the articles or prepositions that precede noun-phrases. Form includes such aspects as frequencies of paragraphs, emphasis tags, images, and links. And content includes such aspects as bags of words, stop-lists, number types, and closed-word sets. Whichever features, or combination of features are used, it is still typical that the whole text is analyzed and subsequently categorized into a single genre.

Whole text approaches tend to be successful because different categories of texts comprise different types and quantities of features. And, to be sure, such approaches have yielded impressive results, finding significant distinctions in categories as diverse as dialect, mode, domain, genre, and author. However, to take some slightly more arcane examples, Miliv and Slane (1994) distinguished narratives from treatises by way of the letters D and S respectively; Gordon (2004) identified the penultimate chapter of Joyce's *Ulysses* by the incidence of the letter C; and Šatava (2006) explains that the Võro-Seto ethnolect differs from standard Estonian by way of the letters Q and D respectively: the nominative plural in Estonian featuring a glottal stop, which is marked by the letter Q in Võro-Seto and the letter D in standard Estonian. Such examples may seem churlish but they serve to demonstrate that distinguishing texts, in and of itself, is not difficult, given enough texts and enough variables (and, presumably, enough researchers).

Our Approach

The possible problems with the approaches listed above are ones of *time* and *compositionality*. With regard to time, search engines operating at *peak* performance can only assess the multiple billions of web documents at the rate of 100 pages per second (Franklin 2008). Assessing whole documents over multiple variables may simply be too computationally expensive; thus, there is a significant trade off between time and accuracy. Of course, technology is constantly improving, and consequently, time may become less of a factor. However, by the same token, it could be equally argued that the expansion of the Internet (around 5 million web sites per month) could easily outpace any advances in technology.

With regard to compositionality, our results from Experiments 1 and 2, suggest that texts are heterogeneous in terms of genres. Indeed, the heterogeneous nature of text is well established (e.g., Kintsch & van Dijk 1978; Mann & Thompson 1988; McCarthy, Briner, Rus, & McNamara 2007; Propp 1968, Teufel & Moens 1991; Swales 1990). And this heterogeneity research extends to multiple-genres within texts (see Bazerman 1995; Crowston & Williams 2000; Orlikowski & Yates 1994). Researchers such as these point to *embedded genres* and *genre systems* wherein a single text may feature multiple genres as in *memos*, which may contain proposals; *trials*, which include examination and cross-ex-

amination; *expository* texts, which may include histories, *narratives*, which may include factual claims, and *blogs*, which may include factual accounts, stories, and reviews.

In our approach, both time and compositionality are considered. However, primarily, we base our approach on our psychological findings on genre recognition from Experiments 1 and 2. Our results from these experiments suggest that humans are able to classify narrative, history, and science genres using as few as the first three words of sentences. The results suggest that sentence-level syntax and word-level frequency features may be sufficient for accurate and reliable genre recognition to occur. In our approach, we present a computational model for genre classification based on these findings. That is, we ask: *can a computational model using less than a third of the words in a sentence accurately classify genre using only word-level and syntactical information*? If such an approach is successful, then issues of time and compositionality can be addressed. Our approach would address issues of time because, feasibly, we could imagine a system that samples just a few sentences (or parts of sentences) from the target text. In requiring such a small sample, computational expense is reduced. Our approach would address issues of compositionality because, feasibly, we could imagine the system returning results as to the genre distribution of the samples. That is, perhaps 80% of the samples are science, 15% history, and 5% narrative. Such a result not only informs us of the main genre of the text, it also indicates potential levels of readability or difficulty of the text.

Of course, bringing the discussion above to fruition requires considerable research. And in Experiment 3, we take just the first step towards our goal. Namely, we analyze the sentences from Experiment 2 to assess what degree of accuracy we can expect when using solely the portion of a sentence that humans require for genre recognition.

Methods

To address our computational question, we conducted a number of basic assessments, suitable for sentence level analysis, using the first *three words*, *five words*, and *whole sentence* for each sentence in the corpus. For the lower bound of sentence fragment length, we selected the conservative size of the first three words of the sentences because this was the lowest average number of words for any of the groups from Experiment 2: (i.e., the lower-skill group: $M = 2.98$ words, $SD = 1.24$). For the upper bound of sentence fragment length, we selected the whole sentence to serve as a baseline.

To conduct our analysis, we used as our dependent variable the genre of the sentences as determined from their original source (narrative, history, science). Our independent (or predictor) variables were calculated using the web-based computational tool, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai 2004) and included *word frequency values* (from the Celex data base, Baayen, Piepenbrock, & van Rijn 1993), *word information values* (from the MRC data base

(Coltheart 1981), and *parts of speech frequency counts* (Charniak 2000). In addition, we also included a *syllable count (www.wordcalc.com)*.

The object of the analysis was to ascertain how well the independent variables (i.e. information similar to that which humans might have available) were able to predict the categories of the sentence fragments. One way of achieving this goal is to conduct a series of *discriminant analyses*. A discriminate analysis is a statistical procedure, culminating with a prediction of group membership (in this case, genre) based on a series of independent variables (in this case, the word and syntax variables mentioned above). To guard against issues of overfitting and colinearity caused by applying multiple predictor variables, we followed established procedures of training and testing the algorithm (see Witten & Frank 2005; McCarthy et al. 2007). Thus, the corpus was randomly divided into a training set (67%) and a test set (33%). Using the training set, we conducted an analysis of variance (ANOVA) to identify and retain only those variables that significantly distinguished the genre groups. We then conducted correlations among these variables and eliminated variables that presented problems of colinearity using $r \geq .70$; the variable with the higher univariate F-value was retained and the lower eliminated. Of the 16 remaining variables, the 14 with the highest univariate F-values were used in a discriminate analysis; there was an item to predictor ratio of 10:1. This procedure was then repeated for data collected from the *five words* the *whole sentence* conditions (see Table 8).

Words	Dependent Variable	Mean Narrative	Mean History	Mean Science	F	η^2
3	Past tenses	177.3 (168.12)	68.18 (136.01)	13.33 (65.98)	20.01	0.23
	Pronoun/ noun phrases	184.04 (167.93)	51.14 (119.51	38.33 (105.42)	17.29	0.20
	Syllables	3.70 (.86)	4.89 (1.46)	4.94 (1.46)	13.21	0.16
5	Reading Grade	1.49 (2.12)	6.40 (3.8)	4.72 (3.57)	29.44	0.30
	Past tense verbs	150.00 (115.61)	60.31 (91.43)	9.52 (43.10)	29.31	0.30
	Pronoun/ noun phrases	141.71 (125.77)	32.54 (96.70)	27.78 (69.03)	19.51	0.22
Whole	Reading Ease	83.06 (15.41)	48.98 (21.78)	51.63 (21.78)	22.43	0.40
	CELEX frequency	2.81 (0.25)	2.34 (0.33)	2.45 (0.31)	15.91	0.32
	Past tense verbs	66.67 (66.83)	62.87 (48.06)	5.85 (19.83)	10.68	0.24

Note: All F-values are significant at $p<.001$; SD appear in parentheses

Table 8: Most significant genre predictor variables for "3 word", "5 word", and "whole sentence."

Having established the predictor variables, we used the *training set* data to generate our discriminant function (the algorithm that calculates the predic-

tion of group membership) and we used those generated predictions on the *test set* data to calculate the accuracy of our analysis. Thus, if the results of the discriminant analysis are statistically significant, then we can claim to have evidence that validates the initial analysis. Such a validation affords application of the model to other text corpora of a similar nature. In this study, as is typical of discriminant analysis studies and as is consistent with previous analyses in this study, the accuracy of the results are reported in terms of recall, precision, and F1.

The results of the discriminant analyses were significant (3-words: χ^2 = 33.689, p <.001; 5-words: χ^2 = 30.127, p <.001; whole sentence: χ^2 = 71.704, p <.001). The accuracy of the models in terms of recall, precision, and F1 were comparable to human results (see Tables 9, 10, and 11). The results suggest that as few as the first three to five words of a sentence contain enough *syntactic* and *word level information* to distinguish between genres.

	Narrative			History			Science		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Test set	0.61	0.74	0.67	0.23	0.33	0.27	0.60	0.38	0.46
All data	0.67	0.73	0.70	0.46	0.52	0.49	0.71	0.59	0.65
Participants	0.85	0.71	0.77	0.71	0.76	0.72	0.68	0.87	0.76

Table 9: "Three word" recall, precision, and F1 results for computational model (test set; all data) compared to participant's performance.

	Narrative			History			Science		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Test set	0.47	0.56	0.51	0.55	0.39	0.46	0.61	0.71	0.66
All data	0.72	0.71	0.72	0.66	0.53	0.59	0.60	0.73	0.66
Participants	0.85	0.71	0.77	0.71	0.76	0.72	0.68	0.87	0.76

Table 10: "Five word" recall, precision, and F1 results for computational model (test set; all data) compared to participant's performance.

	Narrative			History			Science		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Test set	0.70	0.56	0.62	0.62	0.67	0.46	0.60	0.68	0.64
All data	0.75	0.70	0.73	0.64	0.67	0.66	0.75	0.77	0.76
Participants	0.85	0.71	0.77	0.71	0.76	0.72	0.68	0.87	0.76

Table 11: "Whole sentence" recall, precision, and F1 results for computational model (test set; all data) compared to participant's performance.

The three-word model is most impressive at identifying narratives (all data F1 = .70, human = .77) and reasonable at identifying science (all data F1 = .65, human = .76). The three-word model appears weakest at identifying history (all data F1 = .52, human = .72). The five-word model returns similar results although the history identification is improved (all data F1 = .59, human = .72). The whole-sentence model returns human like results for all three genres (narrative: all data F1 = .73, human = .77; history: all data F1 = .66, human = .72; science: all data F1 = .76, human = .76). The results suggest that with some modifications to the model (e.g., genre related frequencies) that a highly accurate subsentential genre identification model is feasible.

Discussion

In Experiment 3, we developed and tested a computational model of human genre recognition at the sub-sentential level. Our results suggest that basic subsentential features such as *parts of speech* and *word frequencies* significantly distinguished between genres. Further, the success of our computational model suggests that the features of only the first three to five words are sufficient for this classification.

The results of our model are particularly impressive when considering humans' advantages when recognizing genre in comparison to our model. For example, the computational models did not contain information about semantics and word knowledge, which humans would likely use when recognizing text genre. Thus, when participants see a number such as 1776 they are presumably more able to interpret this as an historical date. Second, even though word frequency was included as a predictor, the results are based on *frequencies in general* rather than genre specific. We can hypothesize that word information relevant to specific genres would enhance the accuracy of the prediction. For instance, we might assume that participants have knowledge that *cannon* is a word associated with history whereas *nucleus* is a word associated with science. Third, we might further hypothesize that our model could be improved if frequencies were calculated from only the sentence-initial fragment. Thus, words such as *to*, operating as an infinitive marker as in *to understand this process …*, may be more indicative of expository text.

While the results of Experiment 3 suggest that word and syntax variables may be all that humans (and computational models) need to recognize genre, this does not mean that more complex discourse variables such as cohesion variables and temporal features are not a characteristic of genre differences. However, our results do indicate that readers can and do make genre decisions before such features become available. Such a result is important when considering a light and efficient approach to genre categorization where computational expense is an important factor.

Finally, the results of our three-word model are impressive; although we cannot claim that the model is as good as human performance. We have given

modifications above for improving our model, but it is still worth noting that some features of our model do match human performance. For instance, the narrative precision evaluation for the test set (.74), all data (.73), and for participants (.71) are highly similar. Given that Experiment 2 showed that the human narrative precision variable correlated highly with reading skill (r = .520, p = .002), it is reasonable to assume that the computational model might reflect some aspects of reader strategy, at least in its propensity to correctly reject non-narrative decisions for narrative sentences. Additionally, the model's false alarms for narratives were similar to those decisions made by humans: that is, false alarms were less likely to be science decisions (History = 11; Science = 6).

Final Discussion

This study included three experiments designed to address issues concerning genre recognition. Experiments 1 and 2 addressed issues concerning (1) how many words were necessary to constitute human recognition of genre, (2) to what degree were the texts heterogeneous in terms of genre, and (3) to what degree was genre recognition a predictor of reading ability. In Experiment 3, we used the information gathered from the previous two experiments to create a computational model for genre recognition.

Research Questions

Our study began with six research questions. Here, we briefly summarize the responses to those questions based on the current research.

1. *How short (in terms of number of words) can a text be for its genre to be accurately recognized?* Using a baseline of 33%, most readers (77%) can accurately recognize genre within the first three words of a sentence. Many readers (50%) can accurately recognize genre using just the first word. This result suggests that genre is (also) a sub-sentential feature of text.

2. *What types of errors (i.e., genre misclassifications) do readers make when identifying genres?* The most common type of genre misclassification appears to be the assigning of narrative to non-narrative text. We can presume from this result that readers are more familiar with the features of narratives and tend to make a default assumption that a text is narrative unless shown to be otherwise. We can hypothesize that explicit training in recognizing non-narrative features may facilitate reader comprehension if it facilitates earlier and more accurate genre recognition.

3. *To what degree are texts heterogeneous?* Our results suggest that texts are about 83% homogenous in terms in genre. For the remainder, narratives tend to comprise mostly history sentences; histories tend to comprise mostly narrative sentences; and science texts tend to comprise an even number of narrative

and history sentences. We can hypothesize that variation in the heterogeneity of the text may benefit some readers more than others based on their knowledge or skill level. For instance, we can presume that lower skilled/knowledge readers of science texts would be facilitated by a higher incidence of narrative/history sentences because the sentences features used in this genre are likely to be more familiar.

4. *Does the process of genre identification depend on reading skill?* Our results suggest that higher-skilled readers are less likely to think that a sentence is a narrative when it is not, and they also require less time to accurately recognize genre. These results lead us to believe that a reading skill assessment based on genre recognition is viable.

5. *What textual features (e.g., syntax, lexical choice) influence genre identification?* Our results suggest that such features as presence of *past tense*, *length of words*, and *word frequencies* offer readers substantial indication of genre at the sub-sentential level. This result is important for designing and modifying computational approaches to genre classification, as well as forming part of the training for an intervention approach to helping students with reading skills.

6. *Can a computational model categorize genre using only as much text as humans appear to need?* The results of our computational models were statistically significant and comparable to humans. The three-word and the five-word models were most impressive at identifying narratives and science. The whole-sentence model returned human like results for all three genres. We hypothesize that improvements to our word frequency database and using genre-specific word frequencies would significantly improve the computational model. Overall, the results provide a good deal of confidence that computational genre recognition is achievable using only as much sentential information, as is required by humans.

Limitations of our study

While we would argue that the results presented in this study offer a significant contribution to research in genre recognition, the limitations of the study are worth acknowledging. First, having only considered three traditional text genres, we cannot be sure how such analysis would scale up to a finer grained analysis of genres such as those encountered on the internet. In addition, the genres used in the study were presented to participants as their one and only choice. It is possible that participants may have preferred to make multiple choices or categorized sentences in genres other than those we stated. Second, we cannot be sure that the research presented here suitably distinguishes *topic* from *genre*. Addressing this issue is of significant importance to future research. Third, the numbers of sentences and participants in our experiments are rela-

tively small. Such limitations are common in initial forays into new research; however, given such numbers, we must be cautious as to the conclusions we draw. Fourth, while our computational models showed promise, and while our extensions to these models seem reasonable, there is considerable work to be done if we are to establish that such an approach can produce a desirable accuracy while minimizing computational expense.

Conclusion

The research presented here offers an interesting and promising direction toward a better understanding of genre recognition. In psychological terms, we plan to use this research to better establish our *genre identification paradigm* as an assessment of reading skill, and even as a possible intervention for reading development. In computational terms, our results suggest that text classification model at the genre level is possible using only a limited selection of text fragments. Such an approach offers the possibility of fast and accurate genre classification as well as information as to the genre distribution within a text. While much remains to be done, the results presented here offer a new and exciting perspective on the nature of text, the possibilities of new assessments of reading skill, and an intriguing and novel approach to computational text classification.

References

Albrecht, J. E., O'Brien, E. J., Mason, R. A., & Myers, J. L. (1995). The role of perspective in the accessibility of goals during reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21, 364-372.*

Baayen, R. H., R. Piepenbrock, and H. van Rijn (Eds.) (1993). *The CELEX Lexical Database* (CD-ROM). University of Pennsylvania, Philadelphia (PA): Linguistic Data Consortium.

Bazerman, C. (1995). Systems of genres and the enactment of social intentions. In A. Freedman and P. Medway (Eds.), *Genre and the New Rhetoric*. London: Taylor and Francis.

Bhatia, V. (1997). Applied genre analysis and ESP. In T. Miller (Ed.), Functional approaches to written text: Classroom applications. Washington, DC: USIA.

Biber, D. (1988). Variation across speech and writing. New York: Cambridge University Press.

Boese E. (2005). Stereotyping the web: genre classification of web documents, *M.S. Thesis*, Computer Science Department, Colorado State University. Fort Collins, CO.

Boese, E. S. & Howe, A. E. (2005). Effects of Web Document Evolution on Genre Classification. In *proceedings CIKM 05*.

Charniak, E. A Maximum-Entropy-Inspired Parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics* (2000), pp. 132-139.

Clements, P. (1979). The effects of staging on recall from prose. In R.O. Freedle (Ed.) New Directions in Discourse Processing (pp. 297–330). Norwood, NJ: Ablex.

Coleridge, S.T. (1985). Biographia Literaria: Samuel Taylor Coleridge, H.J. Jackson (ed.), Oxford.

Coltheart, M. (1981). The MRC psycholinguistic database quarterly. *Journal of Experimental Psychology, 33A*, 497-505.

Crowston, K. & Williams, M. (2000). Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society* 16, 201-215.

Davies, A, & Elder, C. (2004). *The handbook of applied linguistcs*. Blackwell Publishing. Oxford.

Downs, W. (1998). *Language and society*. Cambridge University Press.

Duran, N.D., McCarthy, P.M., Graesser, A.C., & McNamara, D.S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior, Research and Methods, 29*, 212-223.

Finn A. & Kushmerick N. (2006). Learning to classify documents according to genre.

Journal of the American Society for Information Science and Technology, Special Issue on Computational Analysis of Style, 7.

Franklin, C (2008) http://computer.howstuffworks.com/hsw-contact.htm. Retrieved 03/05/2008

Gerrig, R. (1993). Experiencing narrative worlds: On the psychological activities of reading. Cambridge, MA: MIT Press.

Gordon, J. (2004). *Joyce and Reality: The Empirical Strikes Back*. Syracuse, NY: Syracuse University Press.

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced outlines, familiarity, text genre, and retention of prose. *Journal of Experimental Education, 48*, 209-220.

Graesser, A. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. Journal of Verbal Learning and Verbal Behavior, 19, 131-151.

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36, 193-202.

Graesser, A.C., Olde, B. A., & Klettke, B. (2002). How does the mind construct and represent stories? In M. Green, J. Strange, and T. Brock (Eds.), *Narrative Impact: Social and Cognitive Foundations*. Mahwah: NJ: Erlbaum.

Hatch, E. & Lazarton, A. (1991). Research manual: Design and statistics for applied linguistics. New York: Newbury House.

Hymes, D. (1972). Models of interaction of language and social life. In *Directions of Sociolinguistics: The Ethnography of Communication* (Eds.) J.J. Gumperz & D. Hymes. New York: Holt, Rinehart and Winston.

Karlgren J. and Cutting D. (1994). Recognizing Text Genre with Simple Metrics Using Discriminant Analysis. Proceedings of COLING 1994, Kyoto.

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. Journal of Experimental Psychology: Learning, Memory and Cognition, 29, 439–446.

Kennedy A. & Shepherd M. (2005), Automatic identification of home pages on the

web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In Proceedings of the 35th annual meeting on Association for Computational Linguistics, p.32-38, Madrid, Spain.

Kieras, D. E. (1978). Good and bad structure in simple paragraphs: Effects on apparent theme, reading time, and recall. *Journal of Verbal Learning and Verbal Behavior, 17*, 13-28.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Lee J., Kwong O. Y. (Eds.) Natural Language Processing. Springer, Berlin.

Linderholm, T., & vcan den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. Journal of Educational Psychology, 94(4), 778–784.

Kendou, P. & van den Broek, P. (2005). The effects of readers' misconceptions on comprehension of scientific text. *Journal of Educational Psychology, 97,* 235-245.

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and aging. *Psychology and Aging, 16*, 145-160.

Lee Y. and Myaeng S. (2002). Text Genre Classification with Genre-Revealing and Subject-Revealing Features. *Proceedings of the 25th Annual International ACM SIGIR* : 145-150.

Lee Y. and Myaeng S. (2004). Automatic identification of text genres and their roles in subject-based categorization. *Proceedings of the 37th Hawaii International Conference on System Sciences*.

Lightman, E.J., McCarthy, P.M., Dufty, D.F., & McNamara, D.S. (2007). The structural organization of high school educational Texts. FLAIRS, 2007.

Lim C., Lee K. and Kim G. (2005). Automatic genre detection of web documents. In Su K., Tsujii

Louwerse, M.M., McCarthy, P.M., McNamara, D.S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843-848). Cognitive Science.

McCarthy, P.M., Briner, S.W., Rus, V., & McNamara, D.S. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao, & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 107-122) . London: Springer-Verlag

McCarthy, P.M., Graesser, A.C., & McNamara, D.S. (2006, July). Distinguishing genre using Coh-Metrix indices of cohesion. *Paper presented at the Society for Text and Discourse conference*, Minneapolis, MN

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Duran, N. D., Fujiwara, Y., & McNamara, D. S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British scientists. *Foreign Languages for Specific Purposes*, *6*. 46-77.

McCarthy, P.M., Renner, A.M., Duncan, M.G., Duran, N.D., Lightman, E.J., & McNamara. D.S., (2008). Identifying topic sentencehood. *Behaviorial Research and Methods*, 21, 364-372.

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language, 25*, 645-656.

Meyer zu Eissen S., Stein B. (2004). Genre classification of web pages : User study and feasibility analysis. In Biundo S., Fruhwirth T., Palm G. (Eds.), *Advances in Artificial Intelligence*. Springer, Berlin : 256-269.

Meyer, B. J. F., & Wijekumar, K. (2007). Web-based tutoring of the structure strategy: Theoretical background, deisgn, and findings. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Erlbaum.

Miliv, L. T. & Slane, S. (1994). Qualitative aspects of genre in the century of prose corpus. *Style 24*, 42-57.

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Erlbaum.

Orlikowski, W. J. and Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly*, *33*, 541-574.

Otero, J., Leon, J. A., & Graesser, A. C. (Eds.), (2002). The psychology of science text comprehension. Mahwah, NJ: Erlbaum.

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and. aging. *Psychology and Aging, 16*, 145-160.

Rosso, M. A. (2005). Using genre to improve web search. *PhD. Thesis*. University of North Carolina, Chapel Hill.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik B., Cai, J., & Liu, X. (2001). Genre based navigation on the web. In *Proceedings of the 34th Hawaiian International Conference on System Sciences*, Hawaii. IEEE ComputerPress.

Santini M. (2006). Some issues in Automatic Genre Classification of Web Pages, JADT 2006 - 8èmes Journées internationales d'analyse statistique des données textuelles du 19 au 21 avril 2006 à l'université de Besançon (France).

Šatava, L. (2006). *"Regional languages" as emancipation strategy*. Czech lands in the middle of Europe in the past. *MSM* 0021620827.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G.(2001), Automatic text categorization in terms of genre and author, *Computational Linguistics, 26(4)*, 471–495.

Swales, J. (1990) *Genre Analysis.* Cambridge: Cambridge University Press.

Teufel, S. & Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (Eds.), *Advances in automatic text summarization*, MIT Press, 1999.

Tonjes, M.J., Ray, W., & Zintz, M.V. (1999). Integrated content literacy. New York: The McGraw-Hill Publishers.

Trabasso, T., & Bartolone, J. (2003). Story understanding and counterfactual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 904-923.

Van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Vidal-abarca, E., Martinez, G., & Gilabert, R. (2000). Two procedures to improve instructional text: Effects on memory and learning. *Journal of Educational Psychology, 92,* 107-116.

Williams, P. J. (2007). Literacy in the Curriculum: Integrating Text Structure and Content Area Instruction. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Erlbaum.

Witten, I.H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann Publishers.

Wolfe, M. B. W. (2005) Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 2, 359-364.

Zwaan, R.A. (1993). Aspects of literary comprehension. Amsterdam: John Benjamins.

Appendix

Sample of the sentences used in the study

Index	*Topic sentence value (1-6)*	*Sentence 1 (1)/ Sentence 3 (0)*	*Narrative (0); History (1); Science (2)*	*No. of words*	*Sentence*
1	2.67	0	1	15	Because of the fragmented nature of Mayan society, the different cities frequently went to war.
2	1.33	0	1	11	They moved slowly, not as invading hordes but as small communities.
23	2.00	1	0	18	When the time was up, Mr.Dooley asked us to put down our pencils and pass our tests in.
52	2.00	0	2	14	However, more material is ultimately moved by the greater number of slow mass movements.
53	2.33	0	2	20	Likewise, it's easier to express the concentration of a solution as the number of moles of material dissolved in it.

All sentences used in this study can be downloaded at http://tinyurl.com/55ex2x

Ryan Levering, Michal Cutler

Cost-Sensitive Feature Extraction and Selection in Genre Classification

Automatic genre classification of Web pages is currently young compared to other Web classification tasks. Corpora are just starting to be collected and organized in a systematic way, feature extraction techniques are inconsistent and not well detailed, genres are constantly in dispute, and novel applications have not been implemented. This paper attempts to review and make progress in the area of feature extraction, an area that we believe can benefit all Web page classification, and genre classification in particular. We first present a framework for the extraction of various Web-specific feature groups from distinct data models based on a tree of potentials models and the transformations that create them. Then we introduce the concept of cost-sensitivity to this tree and provide an algorithm for performing wrapper-based feature selection on this tree. Finally, we apply the cost-sensitive feature selection algorithm on two genre corpora and analyze the performance of the classification results.

1 Introduction

A classification task cannot achieve high performance without being provided a good set of features, regardless of the algorithm that is being used to train the computer. For instance, if you were attempting to train a computer to recognize dogs from other animals and the only measurement you took was the number of legs the animal had, it would be an impossible task. This paper is primarily concerned with the goal of obtaining thorough measurements, known as features, from a particular automatic classification domain: Web pages.

The features that we are discussing would most likely be useful for any number of classification tasks within the domain of Web pages. Ideally, there would be a general way, independent of classification task, to measure the effectiveness of a feature in representing the information on a Web page. For instance, when we are dealing with text documents, a bag-of-words model is a fairly good representation of the kind of content on the page. If word order was also included in the model, it would come much closer to representing all the information that is encoded in the document. However, Web pages are very high-dimensional concepts that layer semantic and visual annotation on top of the already rich semantics of text. This makes it much harder to theoretically verify the merit of a feature set. Therefore, the merit is more easily measured with practical experiments in which the features are used.

While we want to achieve a high accuracy with Web page classification, we recognize that practical Web-scale classification implementations cannot afford to compute every possible feature to achieve the maximal accuracy. Even if they could, they should prefer to minimize the time to achieve that accuracy. Therefore, the work presented in this paper focuses on finding a balance between accuracy and efficiency in feature extraction. As an example, if a particular genre class of Web pages is easily recognizable from the URL, it should not be necessary to actually perform a network fetch. While this can be decided manually, we believe that this sort of analysis can be incorporated into the machine learning process to better effect.

We have chosen to evaluate our feature extraction and selection methodology in the context of automatic genre classification tasks. The meaning of *genre* is quite complex and there are many different definitions in the information science literature. For the reader unfamiliar with genre in the context of Web page classification, we advise reading sections of more comprehensive works such as Santini (2007) or Boese (2005). Empirically, common examples of Web page genres include such labels as "FAQ", "News Article", "Company Home Page" that encompass a common perception of a document type.

Genre classification is a good choice to evaluate Web page features because genres of Web pages often need a larger amount of contextual information on the page to make a classification decision and therefore actually require more interesting features. Genre classification, on the machine learning side has always found merit in and been defined by features that traditional topical classification ignored. Stamatatos et al. (2000) found that stop words, which are typically thrown out in topical classification, were good indicators of genre. Toms and Campbell (1999) found that users could identify genre based on visual layout and spacing information. Later in our own research Levering et al. (2008), we corroborated that measuring this visual information could help with certain automatic genre classification tasks. As an example, HTML tag counts are almost always included in feature sets for Web genre classification tasks, when they are frequently stripped out in topical classification.

The general difference in the two types of automatic classification is that the feature spaces of the two types of classification are often orthogonal. From a human perspective, a particular topic will span multiple genres of representation and the genre of a document often does not imply a particular textual content. At the same time, certain genres are in practice correlated with particular topics. Additionally, there are features that are useful in both types of classification. The larger size and diversity of the feature space means that more care and sensitivity needs to be given to feature extraction and selection in genre classification.

This paper will start off in Section 2 with an introduction to a more descriptive methodology of feature extraction on Web pages. Once this method is presented, in Section 3 we will identify several useful data models that we can analyze to obtain Web page measurements. Then in Section 4 we will perform a review of current feature groups used in Web page classification using this more formalized methodology and the models from the previous section. In Section 5 we will introduce a way to use

this extraction methodology to perform a measurement cost-sensitive feature selection across a large general set of feature extraction techniques. In the final section, we will demonstrate both the feature selection algorithm and our set of features on two previously used genre corpora.

2 Feature Framework

In order to compare different methods of classifying Web pages, the extracted features should be comparable. In basic text classification, this is not as great of an issue. Visually interpreted from a character level, text has two obvious forms of abstract representation. First, the visual model that is the concrete level on which authors generally write:

$$\text{glyph} \in \text{line} \in \text{page} \in \text{document}$$

On top of this is the semantic model that actually conveys the meaning of the document. Recognized visual cues (spacing/line breaks/punctuation) from the first textual level are used to produce a more semantic model, for example:

$$\text{character} \in \text{word} \in \text{sentence} \in \text{paragraph} \in \text{section} \in \text{document}$$

This model can produce most of the features that the field is familiar with: bag-of-words, sentence counts, punctuation counts, pattern matches. These features have generally established semantics and procedures for extraction that are agreed upon by a community and thus are comparable across classification experiment.

However, even these accepted features are often used without explanation and can often lead to obfuscated problems and results in classification. Boundary cases, noise, and parameters have to be dealt with in any algorithm implementation. One implementation might throw away any words with less than three letters plus stop words and not count sentences that cannot be parsed (like fragments within ellipses). These details are typically lost during most communication channels, such as papers and emails between researchers. It is also rare to find agreement on a particular software package used to perform feature extraction. While these details often will not make a difference, sometimes they will make analyzing features for classification a frustrating issue.

Another supplementary problem is that this fairly small list of models explodes when you attempt to analyze Web pages. HyperText Markup Language (HTML), the language in which most Web pages are written, is a semantic as well as visual expression language. It allows almost complete flexibility in the order and manner in which text is displayed. There are at least two obvious models that are layered on top of the already existing models for text representation: the HTML model itself, often called the Document Object Model (DOM) in interpreted, hierarchical form and the rendered model which is the way the document is drawn to the screen by the browser for the viewer. We could even go further and say there is a semantic model on top of that visual model that could represent how the author intends the page to be interpreted

by the viewer. The point is that feature extraction in HTML becomes a much harder problem. This is because:

1. There are more models to work with and extract potential information from

2. The transformations between these models are sometimes very complex (in the rendering step for instance) and not agreed upon

3. The models themselves have more dimensions (for instance, graphically two dimensions) in addition to nested textual semantics

For all of the reasons above, we propose that a better framework for comparing feature extraction techniques needs to be established. This paper presents some established and new features we have found useful in the context of this feature framework. We envision feature extraction as a tree of model transformations that starts with a single initial data model and ends with a set of feature groups. In the diagram below, each of the circle nodes represent models that are used to potentially extract features from. The squares represent transformations that convert one type of model to another. The triangles are the actual feature groups that are extracted from each model to perform classification.

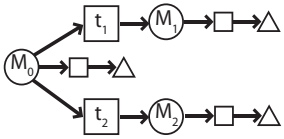


Figure 1: Feature extraction as tree of transformations.

A model is an abstract representation of the original data or a transformation of that data. Thus, the feature groups that result from the transformation steps are also instances of this generic model. A transformation is a higher-level function that takes a set of input models and produces a different set of output models.

These models and transformations are each identified by a unique URI (Uniform Resource Identifier). In the case of a transformation, this URI globally defines the algorithmic behavior and is ideally backed by a specification, or at least a detailed description, identifying working behavior and configuration parameters. In the case of a model, this URI identifies the structure and semantics of the model. This means that each feature in the end feature group is no longer just an identifier, but actually a transformation path that describes specifically how the feature was obtained. The worth of these identifiers are obviously dependent on the level of detail of the specifications or descriptions that define the transformations and models, but at the least they serve as a constrained language for discussing feature extraction.

3 Web Models and Transformations

Presented in Figure 2 is a summarization of the feature extraction tree that we use in our Web page analysis. Not included are all the leaf feature groups that are generated through extraction transformations.

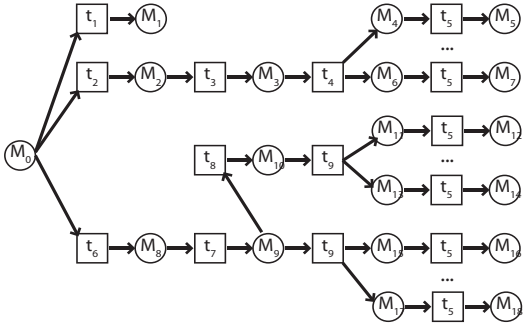


Figure 2: Feature extraction tree to convert an initial URL into models to use in feature extraction.

In order to reuse techniques for feature extraction, we often will create transformations to convert to existing model types. For instance, all the leaf models in Figure 2 are all the same type (fragmented text) even though they represent different textual contents (i.e. extracted URL tokens vs. content tokens).

Many of these transformations deal with two different implementations of the core text extraction path. One of these paths uses a browser-driven rendered DOM version and the other uses a more simplistic DOM version. The general idea is illustrated in Figure 3, which is a simplified view of Figure 2 that collapses similar transformations (listed below the node title) into recognizable steps in a common Web page feature extraction path. We relate some of the underlying details of these models and transformations. Included in them are URI references relative to our own namespace.

Figure 3: Core transformation path.

Web Node (`<model/www-node>`): The model all the way to the left is the Web node model. This is a more realistic form of a URL expanded to include state that is needed for the HTTP request. Any sophisticated general-purpose crawler can possibly track cookie state or issue POST requests that will drastically influence the produced Web

page. Therefore, our model of a Web page location needs to be expanded to deal with this.

Fetcher: We use two different types of fetchers to fetch Web nodes. The first is a modified version of Apache Nutch's fetcher that does single-threaded pure HTML fetching (`<transformation/fetcher/nutch>`). This is very fast and in addition to the standard URL fetching it uses intelligent parsing to handle meta-redirects without a full document parse. The second is a rendered document fetcher that downloads the complete Web page (`<transformation/fetcher/swt>`). This fetcher loads the document using Mozilla's backend rendering engine and then saves the complete page with image, style, and script resources very accurately. By interpreting the page scripts, the HTML output by this fetcher is generally much more accurate than the output produced by the Nutch fetcher.

Web Content (`<model/www-content>`): This is the raw content found at a webnode. It is composed of the resulting URL (which may differ from that requested) and raw bytes, possibly including resources from the fetched URL depending on the fetcher.

Parser: For our lightweight transformation path, we convert the results of a single HTML page into an HTML DOM model using JTidy (`<transformation/generator/tidy>`). This library makes similar heuristic decisions to those made by a browser about how to deal with poorly formed HTML. To generate our rendered HTML model, we use the same rendering fetcher library mentioned above and then build a DOM based on Mozilla's own internal DOM model (`<transformer/generator/swt>`). This includes visual positions used to display the nodes in addition to useful style annotations that we can use to pass on extra information like element visibility.

HTML (`<model/html>`): A parsed HTML document, represented as a hierarchical in-memory Document Object Model.

Rendered HTML (`<model/rendered-html>`): A HTML DOM model annotated with position (height, width, x, y) for each one of the DOM nodes.

Text Stripper: To generate the text, we traverse the DOM model and output text nodes, generating line breaks when certain text-breaking HTML nodes are encountered. The only difference between the rendered HTML text stripper (`<transformer/renderedhtml-text>`) and the basic HTML text stripper (`<transformer/html-text>`) is the extra visibility information that is used so that non-visible text is not converted.

Text (`<model/text>`): Basic string text, decoded bytes interpreted as a character array.

Text Fragmenter: This is a lookahead parser that fragments English text by using punctuation heuristics along with abbreviation lookups (`<transformer/text-ftext>`). It is most likely not as accurate as a heavy-weight statistical parser, but it is very fast and provides much better results then simply splitting on punctuation/delimiters. Web pages often do not have well-formatted sentence structure and it is not uncommon to have Web pages without any complete sentences in the text. Being able to handle fragments as well as common word punctuation (dates/urls/etc) keeps us from losing this extra information when we extract features from the text.

Fragmented Text (<model/ftext>): Heuristically, fragmented text that is broken down into a linear sequence of word and punctuation, grouped by sentence or fragment if the parser did not think it was a valid sentence. This model is much more accurate for token analysis than delimiter split text.

4 Web Feature Extraction Groups

Generally, many features and models have been extracted from Web pages at some point in the literature, though most of them have been understated or implicit. In this section, we hope to highlight explicitly some of the main categories of Web page features in the context of the models they are derived from, along with some examples of how they can be used. We use our formal notation to explain how we represent those features in our system.

4.1 Web Node Feature Groups

Genre classification can often be surprisingly effective with URL driven features. As several very rough examples, a tilde could imply a personal home page, short URL length could imply corporate home page, digit count could imply time-sensitive or version sensitive genres. This is likely more true in genre classification than in topical text classification which is more dependent on a richer topical vocabulary.

We generate simple character statistics on the URL (<extractor/www-content/url>) and then tokenize the URL intelligently (<transformer/www-ftext>) and apply all of our fragmented text feature extraction techniques (see Section 4.3).

4.2 Text Feature Groups

Because most of the features we use are based on our fragmented text model, we do not do much analysis on the raw text. We count characters of different types such as digits, alphanumeric, and punctuation (<extractor/text/simple>) and create a dynamic punctuation vocabulary (<extractor/text/punctuation>).

4.3 Fragmented Text Feature Groups

The core of fragmented text features is the dynamic vocabulary extraction (<extractor/ftext/dynamic-vocab>), which gives us the very common bag-of-words text word counts. Because of the fragmentation, these have already had punctuation dealt with. We do not eliminate stop words because they are often good genre indicators (Stamatatos et al., 2000), but we do perform stemming. Stemming has long been used in text classification as a simple way to combine frequencies of words with similar roots and thus similar semantic concepts. These features are very important to classifier performance as they pick up on genre-sensitive vocabulary.

To add some extra abstraction ability to the classifier, we also include a number of common token patterns such as several different date formats, integer and decimal

numbers, several common time formats, words in upper case, and words in title case (<extractor/ftext/pattern>). Finally we also have some basic fragmentation statistics like sentence counts, word counts, etc. (<extractor/ftext/simple>).

4.4 HTML Feature Groups

The primary HTML features frequently used in Web page have been tag counts (<extractor/html/tag>). These have been mentioned in many classification papers such as Karlgren et al. (1998), Joachims et al. (2001) and Boese's 2005 feature review. Generally, most people at least intuitively agree that high tag counts can be indicative of the type of a Web page. For instance, heavy TABLE (a layout annotation) tag usage may indicate complex layout or many IMGs (representing page images) may indicate a visually complex page and hubpages are essentially defined by having many A (hypertext link) tags. They can also be useful for discovering special cases of Web pages that rely on the presence of a more specific tag (like a file upload tag).

To these basic features we add several sets of slightly more interesting abstractions. Form elements get their own treatment, so we count the number of checkboxes, radio buttons, etc (<extractor/html/form>). These features can point to interactive genres (contact, feedback). We also count table and HTML depth that can be used to detect complex page layouts (<extractor/html/depth>). Counts of event handlers on tags (<extractor/html/events>) can point to more interactive, JavaScript-laden genres. Finally, we analyze the link tags to see what type of file and what domain they are pointing to (<extractor/html/link>). This is especially useful for a download or link type genre.

4.5 Rendered HTML Feature Groups

One advantage of a rendered DOM comes when you use it to "clean" standard DOM features. The DOM produced is structurally the same as a non-annotated DOM so it can be used in the same way. For instance, often there are parts of a Web page that are dynamically generated or that are listed in the HTML source but are not actually visible on the page. With rendering information, you can very easily tell that these elements are not included and thus have a potentially cleaner version of the page contents. This especially applies to site home pages, where dynamic visual updating often plays a large part of the user experience. Most features generated without rendering are essentially guesses or approximations of what is actually going on when the page is rendered.

Other features can be generated by examining overall statistics of the extra visual positions. These were discussed in Levering et al. (2008) in the context of genre classification as being useful for identifying certain genre that were dependent on content-type statistics (like containing an image heavy header followed by a lot of text).

Finally, we use the rendered HTML to generate subtree models of the HTML that occur in important parts of the visual document (<transformer/html-section>). This allows us to apply all of the same feature extraction techniques to generate location-aware feature counts. For instance, it allows the classifier to have such information as "a date

appears near the top of the document". In this paper, we use this to analyze the center of the page as in Kovacevic et al. (2002). Because these are fairly specific features, they lend themselves to finer grained genre classification tasks and tend not to show up in feature extraction on broad genres.

5 Cost-Sensitive Feature Analysis

It is impossible to talk about many of these more complex features without also talking about their measurement cost. Any implementation of a genre classification system applied to the Web has to be efficient and most likely scalable to a large number of documents. Luckily, viewing feature extraction as a transformation tree lends itself very well to parallel processing. Each transformation is a discrete functional process and can be batched for large-scale runs in a distributed architectural paradigm like MapReduce (Dean and Ghemawat, 2008).

There is often a diminishing return on the calculation of more complex features, particularly for certain fine grained genres with dominant features. For instance, if a genre or set of genres in a multiclass problem were detectable to some level of accuracy from the URLs and it cost ten times as much computational power to calculate rendered HTML features that would improve accuracy a small percentage, it may not be worth it to calculate those features for a one million document universe let alone the entire Web.

This measurement cost can be factored into the classification process itself using techniques from other classification areas where measurement cost is even more crucial. In Paclík et al. (2002), they use a greedy wrapper-based algorithm to do feature selection based on classifier performance with different groups of features that share computation cost. This was designed for independent feature group measurement costs in the image classification domain, but can be adapted to our Web feature extraction tree paradigm fairly well.

In the following paragraphs, we present an algorithm that performs cost-sensitive feature selection on a feature extraction tree. We chose to simplify the analysis by assuming that once a model is selected all its features can be added at zero cost. We also assume that all these features are presented to the classification algorithms (which may or may not perform additional feature selection). We plan to expand the algorithm shown here to work with more general transformation graphs and to include also the cost of feature extraction for some more costly features.

We assume also that each transformation t of the feature extraction tree has a cost c_t. This is the average computation time required by the transformation t of a model to a child model in the tree. The algorithm depends on a performance evaluation function, $Eval$, that evaluates the performance of a classification task when all features of a set of models S, $features(S)$, are presented to it. The first model included in S is the WebNode model M_0. The goal of the algorithm is to return a set of models S, with a low cumulative computation cost, whose features will produce a performance $P = Eval(features(S)) \geq P_{goal}$.

Let $adjacent(S)$ be the set of models that are not yet in S and are adjacent to at least one model in S. For each model $M \in adjacent(S)$ we can compute the performance improvement per unit of cost achieved by adding M to S. Let $S' = S \cup M$, and let P' be the performance measure of S', the performance improvement $R_{S'} = (P' - P)/c_t$. The next model added to S by the algorithm is the model $M_{max} = argmax\{R_{S'} | M \in adjacent(S)$ and $S' = S \cup M\}$. Using a more complex ratio function or a more complex evaluation of the cost would allow for a more customized feature selection.

Algorithm 1 Greedy feature extraction tree search

```
1: S ← {}
2: P = Eval(features(S))
3: while P < P_goal do
4:    R_max = −∞, P_max = 0, S_max = {}
5:    // Select best model in adjacent(S)
6:    for all M in adjacent(S) do
7:       S' = S ∪ M
8:       P' = Eval(features(S'))
9:       if R_S' > R_max then
10:         R_max = R_S', P_max = P', S_max = S'
11:      end if
12:      S = S_max, P = P_max
13:   end for
14: end while
```

Another possibility with a higher analysis cost is to flatten the feature extraction tree by consolidating all ancestor models into a single model set that contains the cumulative models up to a certain point in a tree. Then we could evaluate every node path at the same time and iteratively choose the one with the best performance gain to measurement cost ratio. This approach would translate to the more greedy previous algorithm if you evaluated any descendant path as opposed to just an adjacent edge. This would also translate the problem more closely into the type of problem in Paclík et al. (2002).

Finally, at the far end of the complexity of analysis scale, an exhaustive search of every combination of feature paths could be done to guarantee the best choice of models for the highest classification performance to cost ratio. If the classification task was fully offline (not recalculated frequently) or the tree was very simple, this would be the best option. For our purposes, this search was not practical due to training cost.

The first two techniques do have the caveat that they depend on classifiers that are resistant to noisy, useless features. Otherwise, greedy choices may not find an optimal solution when an intermediate transformation produces poor features. Paclík et al. (2002) approached this by proposing a nested feature subset selection step while

evaluating a feature group. We use SVMs for evaluation, which we find in practice are very good at ignoring useless features.

6 Experimental Results

We chose to evaluate our feature extraction tree on two datasets to emphasize several different outcomes of feature selection. We used the same transformation costs on both of the datasets. In reality, the transformation costs are dataset dependent, but the magnitudes tend to be fairly consistent. In addition, our corpora are cached HTML pages and so getting accurate fetch times was difficult. In addition, one of the datasets does not include dependent resources (scripts, styles, and images) and therefore accurate renderings were not possible.

We use the greedy tree search discussed above to determine the order of models to generate for feature extraction. This produces a graph with an increasing performance over cost. By setting the P_{goal} to 1.0, we can see the maximum performance of the search algorithm.

6.1 KI-04 Dataset

The KI-04 dataset presented in Eissen and Stein (2004) is a universal set of Web super-genres that can theoretically encompass any Web page. Though it does suffer from some granularity issues (Santini, 2006), it is one of the most complete genre corpora. The lack of a large negative class makes it a genre palette where a multi-class classification makes sense.

It is composed of eight genres: download, article, help, FAQ, private portrayal, non-private portrayal, shop, and linked list. It has over one hundred of each genre type, a palette that was created via user survey.

To evaluate both the feature selection algorithm and the worth of various features on the dataset, we use a ten-fold cross-validating multiclass linear SVM as our *Eval* function. This is the Weka (Witten and Frank, 2005) implementation that uses a pair-wise voting scheme to choose a single class after comparing the results of multiple binary SVMs. At several points, we also ran a multiclass J4.8 classification to validate our results manually by examining the generated decision tree.

Performance in this experiment was measured using accuracy, or number of correct classifications divided by the number of total classifications. This was just to make it consistent with the KI-04 results. However, in a multiclass problem with no large negative class, the accuracy results tend to be consistent with more balanced metrics such as F_1, which we use in the next experiment.

The results are presented in Figure 4. The graph highlights the tradeoffs in classification performance with increasing computation cost. Each data point in the graph corresponds to the cost and performance (in F_1-measure) of a particular feature set that the feature selection algorithm selected as it traversed the cost-weighted feature transformation tree. A labeled version of that tree, showing the evaluated performance

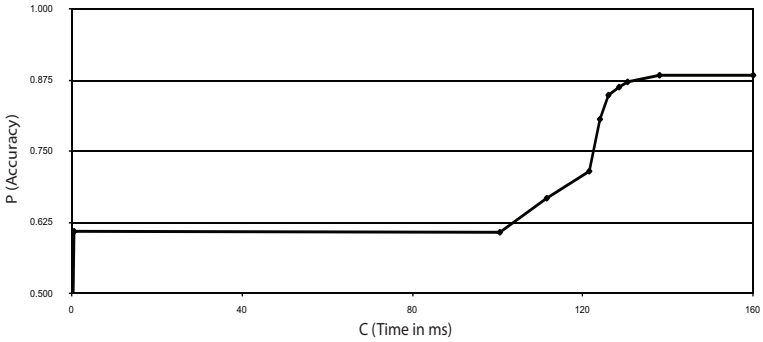


Figure 4: KI-04 feature extraction performance vs. costs.

as we include additional models is shown in Figure 5. Those performance numbers are obtained by measuring the classification performance of the union of the labeled model with every model with a lesser performance value. Also note that having continuously incrementing performance values by adding models is not common. Often there is no gain by including additional models and sometimes they will actually lower the performance, even with noise-resistant classifiers like SVMs.

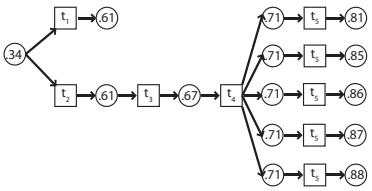


Figure 5: KI-04 feature extraction performance graph.

As an example of that, we did not include any rendered DOM-driven features including visual distributions in this result as their cost-performance ratio was very poor on this corpus and they were added at the very end for no gain. This is due to the previously mentioned point that these features do not work well on very broad non-visual genres in addition to the fact that the dataset was not collected with external resources (scripts, styles, and images), so the renderings were not always very complete.

The long line that stretches most of the graph is the fetch time. One interesting point is that we were able to achieve over sixty-percent accuracy with only the URL. A combination of URL vocabulary like 'FAQ', 'thread' and 'download' combined with URL path lengths did a moderately successful job of classification without any of the overhead of actually fetching the page. If we had trained binary classifiers to just

distinguish FAQ or download genres in this corpus, these results would have been much higher.

After the fetch time, which clearly dominates the computation time, the algorithm did not have too many choices in our current graph without the useless visual features. It added HTML features, then text statistics, then it chose between fragmenting different extracted text types (headings, links, title, emphasized, and all). Interestingly, it found the tradeoff worst on the *all text* category and had better luck with categories like link and heading text. This suggests looking at particular sets of annotated text as opposed to entire vocabulary lists in this type of genre classification.

The research in which this dataset was first presented achieved an overall 70% accuracy on the entire dataset. Thus, our ideal feature set with a maximum F_1-measure of close to 0.9 shows a dramatic overall improvement on the corpus. We theorize that the primary reason for this is the inclusion of URL-based features that appear to have not been used in their analysis.

6.2 Retail Store Dataset

To contrast with several aspects of the KI-04 dataset, we also performed an evaluation on the retail store dataset that we used in previous research (Levering et al., 2008). This dataset has a genre palette fully within the retail store universe of Web pages. It has three positive, labeled genres - store home page, store product list, and store product page - with over a hundred examples of each. It also has a large negative class of other pages in the retail store Web page universe. The goal is to train binary classifiers to be able to recognize those types of pages out of the noise of the entire Web site.

In that paper, we concluded that visual features were useful for this particular problem and raised the accuracy of classification by a significant amount. We wanted to revisit that now with measurement cost factored in to figure out what that gain is costing.

Performance in this experiment was measured as F_1-measure. This is a standard classification metric that balances precision and recall. Generally, there is an inverse relationship between these two metrics, since you can always overtrain a classifier to improve precision at the cost of recall. F_1 provides a single metric that penalizes either measurement being poor.

We evaluated the dataset using a ten-fold cross-validating linear SVM like for the previous experiment but with a single classifier for each of the three positive classes against every other class. This way, we were able to generate a separate performance-cost graph for each classification. The graphs are shown in Figure 6.

One of the most dramatic things about this graph is, like in the previous dataset, just how well a URL alone predicts genre with an extremely low computation cost. You could never use a system with an F_1-measure around 0.5 as shown in the figure for *store-products*, but it does show the feature group's worth. On the other hand, it was intuitively obvious that *store-fronts* were easy to figure out from the URL and the classifier agreed with a nearly perfect F_1-measure.

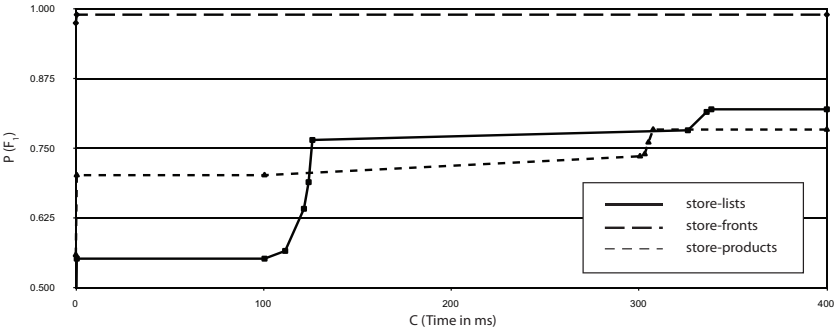


Figure 6: Retail store feature extraction performance vs. costs.

The *store-lists* genre favored non-rendered textual models until it stopped getting improvements, whereupon the search algorithm finally paid the performance cost to render the page in exchange for slightly increased performance. It then produced several more inexpensive models for visually central feature groups that were slightly more accurate than the general feature groups.

Our *store-products* genre did not perform well on non-rendered HTML features (the F_1-measure actually decreased), so the search algorithm opted to render the page and then had several more textual feature groups that it found useful. This is a case of the algorithm choosing incorrectly. If the algorithm included a look-ahead or the second approach was used, it most likely could have found gains with textual features after generating the HTML model without paying the rendering cost.

In all, this experiment verified that while visual features did improve performance on this classification task, simpler features could get nearly the same level of performance. We theorize that it could be useful to have a multi-tiered classifier that first extracts lightweight features. If these features strongly indicate a certain genre, then we can stop extraction; otherwise, we extract more complex features. This is left to future work.

7 Conclusion

The goal of this research was neither to point out interesting facets of the experimental datasets nor to show improvement on classification tasks. Yet at the same time, we made some gains on both of these fronts. Both of our experiments showed that genre is particularly sensitive to URL features and some effort should be spent on more interesting tokenizations and patterns of URLs. Web pages are not just text documents and in a classification task every bit of relevant information should be used.

We improved the accuracy on the KI-04 dataset by using some dynamic URL tokens and textual vocabulary. More importantly, it was done without any focused effort. The same process was applied on both classification tasks to extract features. By having a process to analyze a new classification task on a very generic and powerful feature extraction platform and then perform a cost-sensitive feature selection, we insure that we get acceptable performance while not using unnecessary transformation or extraction techniques.

The real goal of this research was to attempt to break down the process of Web page feature extraction into smaller functional units (transformations) that could be more easily compared and analyzed. We proposed a tree-based abstraction for feature extraction where intermediate models can have their feature groups extracted. Transformations between these models are represented by URIs that convey the semantics of a particular transformation.

One of the benefits of this feature extraction tree is that we can perform a cost-sensitive feature selection as described in this paper. Having intermediate models also would allow researchers to more easily build complex features without repeating earlier work. Finally, by using methods backed by URIs that represent certain algorithmic choices, researchers can more quickly and accurately communicate results.

This work has many directions for improvement. A methodology without a concrete tool that supports the methodology is quickly forgotten. A tool or library to allow sharing of common transformation and extraction techniques would be very productive, much in the same way that Weka has improved the ease of classification research.

Also, there are always more interesting features to be discovered. Visual features may not improve performance on all genre classification tasks, but the Web is becoming a more visual, media-intensive environment and Web classification researchers need to find ways to use these extra layers of information.

References

Boese, E. (2005). Stereotyping the web: Genre classification of web documents. Master's thesis, Colorado State University.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113.

Eissen, S. M. and Stein, B. (2004). Genre classification of web pages: user study and feasibility analysis. *KI-2004: Advances in Artificial Intelligence*, pages 256–269.

Joachims, T., Cristianini, N., and Shawe-Taylor, J. (2001). Composite kernels for hypertext categorisation. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 250–257, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., and Wolkhert, N. (1998). Iterative information retrieval using fast clustering and usage-specific genres. In *Eighth DELOS Workshop - User Interface in Digital Libraries*, pages 85–92.

Kovacevic, M., Diligenti, M., Gori, M., and Milutinovic, V. (2002). Recognition of common areas in a web page using visual information: a possible application in a page classification. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*, page 250, Washington, DC, USA. IEEE Computer Society.

Levering, R., Cutler, M., and Yu, L. (2008). Using visual features for fine-grained genre classification of web pages. In *HICSS '08: Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 131, Washington, DC, USA. IEEE Computer Society.

Paclík, P., Duin, R. P. W., Kempen, G. M. P. v., and Kohlus, R. (2002). On feature selection with measurement cost and grouped features. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 461–469, London, UK. Springer-Verlag.

Santini, M. (2006). Common criteria for genre classification: Annotation and granularity. In *Proceedings of the workshop on text-based information retrieval*.

Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814, Morristown, NJ, USA. Association for Computational Linguistics.

Toms, E. G. and Campbell, D. G. (1999). Genre as interface metaphor: Exploiting form and function in digital environments. In *HICSS '99: Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Chaker Jebari

A New Centroid-based Approach for Genre Categorization of Web Pages

In this paper we propose a new centroid-based approach for genre categorization of web pages. Our approach constructs genre centroids using a set of genre-labeled web pages, called training web pages. The obtained centroids will be used to classify new web pages. The aim of our approach is to provide a flexible, incremental, refined and combined categorization, which is more suitable for automatic web genre identification. Our approach is flexible because it assigns a web page to all predefined genres with a confidence score; it is incremental because it classifies web pages one by one; it is refined because each web page either refines the centroids or is discarded as noisy page; finally, our approach combines three different feature sets, i.e. URL addresses, logical structure and hypertext structure. The experiments conducted on two known corpora show that our approach is very fast and outperforms other approaches.

1 Introduction

As the World Wide Web continues to grow exponentially, web page categorization becomes increasingly important in web searching. Web page categorization, called also web page classification, assigns a web page to one or more predefined categories. According to the type of category, categorization can be divided into sub-problems: topic categorization, sentiment categorization, genre categorization, and so on.

Recently, more attention has been given to automatic genre identification of web pages because it can be used to improve the quality of web search results — see, for example, all the articles in this journal and Mehler et al. (2009).

However, although potentially useful, the concept of "genre" is difficult to define and genre definitions abound. Generally speaking, a genre is a category of artistic, musical, or literary composition characterized by a particular style, form, or content [1], but more specialized characterizations have been proposed. For instance, Kessler et al. (1997) defined a genre as a bundle of facets, focusing on different textual properties such as brow, narrative and genre. According to Shepherd and Watters (1998), while non-digital genre is defined by the tuple <content, form>, the genre of web pages (or "cybergenre") is characterized by the triple <content, form, functionality>, where the functionality attribute accounts for the interaction between the user and the web page. Rauber and Müller-Kögler (2001) defined the genre as a group of documents that share the same stylistic properties. In their experiment with digital libraries, documents of

[1] For example, see Merriam-Webster Online Dictionary `http://www.m-w.com`

the same genre are rendered with the same color. According to Finn (2002), genre is orthogonal to topic, and relates to polarities such as subjectivity/objectivity and positivity/negativity. For Boese (2005) a genre is characterized by the same style, form and content.

In this article, the word "genre" is loosely defined as a textual category that can be more or less related to the topic or content of a web pages. For this reason, we use two different collections. One created by genre researchers for whom the concept of genre is independent from topic (the KI-04 corpus, see Section 4); the other including a number of academic categories (the WebKB collection, see Section 4).

In order to comply with our view of genre, our approach is flexible, incremental, refinable and combines different feature sets. We devised it to be fast, so that in the future it can be applied to web search engines.

Currently, search engines use keywords to classify web pages. Returned web pages are ranked and displayed to the user, who is often not satisfied with the result. For example, searching for the keyword "machine learning" will provide a list of web pages containing the words "machine" and "learning". These pages are from different genres. Therefore, web page genre categorization could be used to improve the retrieval quality of search engines (e.g. see Meyer Zu Eissen (2007)). For instance, a classifier could be trained on existing web directories and be applied to new web pages. At query time the user could be asked to specify one or more desired genres so that the search engine would returns a list of genres under which the pages would fall.

However, a web page is a complex object that includes heterogeneous elements with different communicative purposes. Generally, a web page is composed of different sections organized in the form of headings and links. These sections belong to different genres. Graphical elements (search buttons, images, menus, forms, etc.) and text types, sizes and colors are used to mark sections in web pages. Our approach assigns a web page to all predefined genres with different confidence scores, which represent the similarity between the web page and the centroid of each genre.

It is worth noting that web genres evolve over time because of the continuous modification of the content and purpose of web pages. Simply put, web genre evolution consists in updating old genres and creating new ones. In our approach we focus on the adjustment of old genres. Since automatic genre identification of web pages requires continuous learning, because web pages are often updated, we propose an incremental approach (see Section 3).

Additionally, the Word Wide Web is an open environment, where the user can add a new page, modify the content of actual web page, delete a web page and so on. For this reason, the web is instable and contains many noisy web pages. Taking such web pages into account decreases the accuracy of genre classification (e.g. see Shepherd et al. (2004)). In our approach we propose a refined genre classification of web pages to discard noisy web pages. A web page is considered noisy where its similarities to all genre centroids are below a predetermined threshold.

As mentioned above, a web page is not only a text but contains many HTML tags. The information delimited by these tags is very useful for genre categorization. These

information sources are heterogeneous because they have different representation structures that should be combined to increase the performance of genre classification of web pages.

In summary, the aim of our approach is to provide a flexible, incremental, refined and combined categorization, which is more suitable for automatic web genre identification. Our approach is flexible because it assigns a web page to all predefined genres with a confidence score; it is incremental because it classifies web pages one by one; it is refined because each web page either refines the centroids or is discarded as noisy page; finally, our approach combines three different feature sets, i.e. URL addresses, logical structure and hypertext structure.

This article is organized as follows: in Section 2 we summarize previous work on genre categorization of web pages; in Section 3 we explains our approach; in Section 4 we briefly describe the corpora used in our experiments; Section 5 presents our experimental results; Section 6 presents a comparative study; finally, in Section 7 we present some conclusions as well as our future work.

2 Related Work

Previous work on automatic genre identification is reviewed by focusing on features, classification algorithms and genre corpora.

Features Many types of features have been proposed for automatic genre categorization. In the following paragraphs, the most important features are listed.

Kessler et al. (1997) used four types of features to classify part of the Brown corpus[2] by multiple facets (i.e. brow, narrative and genre). The first type is represented by structural features, which include counts of functional words, sentences, etc. The second type relies on lexical features, which include the presence of specific words or symbols. The third kind of features are character level features, such as punctuation marks. The fourh kind of features is based on derivative features, which are derived from character level and lexical features. These four features sets can be divided into two coarser types: structural features and surface features.

Karlgren (1999) used twenty features including frequencies of functional words and Parts-of-Speech (POSs). He also used text statistics, e.g. counts of characters, words, number of words per sentence, etc.

Stamatatos et al. (2000) identified genre using the most English common words. They used the fifty most frequent words on the BNC corpus[3] and the eight most frequent punctuation marks (period, comma, colon, semicolon, quotes, parenthesis, question mark and hyphen).

Dewdney et al. (2001) adopted two features sets: BOW (Bag Of Words) and presentation features. Presentation features amounted to 89 features including layout features,

[2] http://en.wikipedia.org/wiki/Brown_Corpus

[3] http://www.natcorp.ox.ac.uk

linguistic features, verb tenses, etc. Finn and Kushmerick (2003) used a total of 152 features to differentiate between subjective vs objective news articles and positive vs negative movie reviews. Most of these features were represented by the frequencies of genre-specific words. Meyer Zu Eissen and Stein (2004) used different kinds of features including presentation features (i.e. HTML tag frequencies), classes of words (names, dates, etc.), frequencies of punctuation marks and POS tags. Kennedy and Shepherd (2005) used a feature set including features about content (e.g. common words, met tags), about the form (e.g. number of images) and about the functionality (e.g. number of links, JavaScripts). Boese and Howe (2005) used different kind of features, which can be grouped into three classes, namely stylistic features, form features and content features. More recently, Santini (2007) and Lim et al. (2005) tried to exploit all previously used features. Additionally, Lim et al. (2005) used the URL as new feature and Kanaris and Stamatatos (2007) used character n-grams extracted from both text and structure. Mehler et al. (2007) studied the usefulness of logical document structure in text type classification. They adopted two approaches, which are the Quantitative Structure Analysis (QSA) and the Document Object Model Tree Kernel (DomTK). They conducted experiments to stress the usefulness of structure in document type recognition and compared the QSA approach against the DomTK approach.

Machine Learning Techniques Once a set of features has been obtained, it is necessary to choose a categorization algorithm. Most genre categorization algorithms are based on machine learning (cf. Mitchell (1997)) techniques . Among these techniques, we briefly explain Naïve Bayes, k-Nearest Neighbor, Decision trees and Support Vector Machine techniques because they have been widely used in automatic genre identification.

Naïve Bayes is a simple probability algorithm that determines the probability of a document to belong to a particular genre. Naïve Bayes is a very fast learning algorithm, which is robust to irrelevant features. It needs reduced storage space and can handle missing values. However, since the weights are the same for all features, performance can be degraded by having many irrelevant features. This technique has been implemented by Argamon et al. (1998); Dewdney et al. (2001); Santini (2007).[4]

The k-Nearest Neighbor (k-NN) algorithm groups documents within a vector space. The Term Frequency Inverse Document Frequency (*tfidf*) is usually employed to represent documents. The similarity between documents is computed with Euclidean or cosine measures. New documents are classified with the same genre as the nearest neighbor. The K represents how many neighbors should be analyzed. K-Nearest Neighbor is used only by Lim et al. (2005).

Decision trees are a popular technique used by Argamon et al. (1998), Dewdney et al. (2001) and Finn (2002). Interestingly, Karlgren (1999) applied a combination of decision trees and Nearest Neighbor. He calculated textual features for each document and categorized them into a hierarchy of clusters based on C4.5 *if-then* rules. The

[4]Santini (2007) tried out also Naïve Bayes with different weights.

labels for genres were then decided using Nearest Neighbor assignments and cluster centroids.

Support Vector Machine is a powerful learning method introduced by Vapnik (1995) and successfully applied to text categorization by Joachims (1998). SVM is based on Structural Risk Maximization theory, which aims to minimize the generalization error instead of relying on the empirical error on training data alone. The Support Vector Machine technique has been used in genre categorization by many authors (e.g. Kanaris and Stamatatos 2007; Dewdney et al. 2001; Meyer Zu Eissen and Stein 2004; Santini 2007).

Corpora and Evaluation To date, web genre benchmarks built with principled and shared criteria are still missing (cf. Santini and Sharoff in this issue). This means that the performance of a genre categorization system depends on the specific corpora being classified. For instance, Kessler et al. (1997) used a corpus of 499 texts from Brown Corpus belonging to six diverse genres (reportage, scientific and technical, fiction, etc). They report 0.61 and 0.75 accuracies for logistic regression and neural network classifiers, respectively. Dewdney et al. (2001) used a corpus of 9705 texts belonging to seven diverse genres (advertisements, bulletin, boards, radio news, etc.). They achieved 0.83, 0.88 and 0.92 accuracies for Naïve Bayes, C4.5 and SVM, respectively. Meyer Zu Eissen and Stein (2004) compiled the KI-04 corpus. In their first experiment, they used 800 web pages (100 web pages for each of the eight genres included in the corpus) and applied discriminant analysis. They achieved an accuracy of 0.7. Boese and Howe (2005) used the WebKB corpus to study the effect of web genre evolution. Based on logistic regression classifier, they reported an accuracy of 0.8. Kanaris and Stamatatos (2007) used the KI-04 corpus and the SVM classifier. They obtained accuracy between 0.90 and 0.95. Santini (2007) used SVM to classify the KI-04 corpus. She reported an accuracy of about 0.7. Mehler et al. (2007) used SVM classifiers and a German newspaper corpus that contains 31,250 texts distributed over 31 genres or types. Their experiments provided F1=0.78 for QSA and F1=0.57 for DomTK.

3 Proposed Approach

The aim of our approach is to classify web pages by genre based on three different feature sets, namely URL addresses, logical structure and hypertext structure. The proposed approach is based on the construction of genre centroids using a set of genre-labeled web pages. Each new web page is assigned to all genres with different confidence scores, which represent the similarity between the web page and the centroid of each genre.

In the subsection 3.1, we explain our feature extraction process. The representation of features, the construction of centroids, the categorization of new web pages and the combination of classifiers are described in subsections 3.2, 3.3, 3.4 and 3.5, respectively.

3.1 Feature Extraction

In our approach, we used three different types of features, which are the URL addresses, the logical structure and the hypertext structure.

The URL is encoded as a text line, which contains genre-specific words. For example, the presence of "FAQ" and "CV" in the file name is a reliable hint of the membership of a web page to the FAQ and CV genres, respectively.

The logical and hypertext structures of a web page are encoded into the HTML tags used in the web page. The logical structure is represented by the text between $< title >$ and $< /title >$ tags and the text between $< Hn >$ and $< /Hn >$ tags $(n = 1, \ldots, 6)$, while the hypertext structure is represented by the text included in the anchors (between $< A \ldots >$ and $< /A \ldots >$ tags).

To quantify the contextual and structural information, we used the bag-of-words approach – already employed by (Dewdney et al., 2001) for automatic genre identification) – which relies on all words without ordering.

3.2 Representation

Web page representation is performed through three main steps, which are pre-processing, term weighting, and normalization.

Pre-processing Pre-processing is a basic step in document categorization. In our approach, the aim of this step is summarized in the following points:

- Tokenize text into words.
- Remove numbers, non-letter characters and special characters.
- Remove stop words, which are automatically identified using the Luhn Law (Luhn, 1958).
- Use the information gain to reduce the number of obtained terms (Yang and Pedersen, 1997).
- Stem selected terms using the Porter stemmer (Porter, 1980).

Term weighting In our work, web pages are represented using the vector space model. We use three different vectors representing the URLs, the logical structure and the hypertext structure. For each feature set, a web page is represented by a vector p_j of terms. Each term t_i is weighted using the $tfidf$ weighting technique (Salton and Buckley, 1988).

With this technique, the w_{ij} of a term t_i in a web page p_j increases with the number of times that the term t_i occurs in the page p_j and decreases with the number of times the term t_i occurs in the collection. This means that the importance of a term in a page is proportional to the number of times that the term appears in the page, while

the importance of the term is inversely proportional to the number of times that the term appears in the entire collection. Formally, this reasoning is defined as follows:

$$w_{ij} = \frac{tf_{ij}}{max(tf_{1j})} \times \log(\frac{|D|}{n_{t_i}}) \quad (1)$$

where tf_{ij} is the number of times that term t_i appears in web page p_j, $|D|$ is the total number of pages in the collection, and n_{t_i} is the number of pages where term t_i appears.

Normalization The $tfidf$ technique favors large documents and penilizes short documents. To deal with this problem, Lertnattee and Theeramunkong (2004) proposed a normalization technique, called TD, which based on term distribution within a particular class and within a collection of documents.
The term distribution is based on three different factors. These factors depend on the average frequency of the term t_i in all pages of genre g_k. This average, denoted by $\overline{tf_{ik}}$ is defined as follows:

$$\overline{tf_{ik}} = \frac{\sum_{p_j \in g_k} tf_{ijk}}{|D_{g_k}|} \quad (2)$$

where D_{g_k} represents the set of web pages that belongs to genre g_k, and tf_{ijk} is the frequency of term t_i in page p_j of genre g_k.

The normalization factors are the interclass standard deviation ($icsd$), the class standard deviation (csd) and the standard deviation (sd).
The inter-class standard deviation promotes a term that exists in almost all genres but its frequencies for those genres are quite different. For a term t_i, this factor is defined as follows:

$$icsd_i = \sqrt{\frac{\sum_k [\overline{tf_{ik}} - \frac{\sum_k \overline{tf_{ik}}}{|G|}]^2}{|G|}} \quad (3)$$

The class standard deviation of a term t_i in a genre g_k depends on the different frequencies of the term in the pages of that genre, and varies from genre to genre. This factor is defined as follows:

$$csd_{ik} = \sqrt{\frac{\sum_{d_j \in g_k} [tf_{ijk} - \overline{tf_{ik}}]^2}{|g_k|}} \tag{4}$$

The standard deviation of a term t_i depends on the frequency of that term in the pages in the collection and is independent of genres. It is defined as follows:

$$sd_i = \sqrt{\frac{\sum_k \sum_{d_j \in g_k} [tf_{ijk} - \frac{\sum_k \sum_{d_j \in g_k} tf_{ijk}}{\sum_k |g_k|}]^2}{\sum_k |g_k|}} \tag{5}$$

Using the $tfidf$ weighting technique and term distributions for normalization, the weight of term t_i for page p_j in genre g_k is defined as follows:

$$wtd_{ijk} = w_{ij} \times sd_i^{\alpha} \times icsd_i^{\beta} \times csd_{ik}^{\gamma} \tag{6}$$

where α, β and γ are the normalization parameters, which were used to adjust the relative weight of each factor and to indicate whether they were used as a multiplier or as a divisor for the term's $tfidf$ weight, w_{ij}. An experimental study is conducted in section 4 to identify the appropriate values of these parameters.

3.3 Construction of genre centroids

The centroid of a particular genre g_j is represented by a vector G_j. This centroid is the combination of the vectors p_j belonging (or not) to that genre. Several ways were proposed to calculate this centroid. The most used one is the normalized sum, defined as follows:

$$G_j = \frac{1}{\|g_j\|} \cdot \sum_{p_i \in g_j} p_i \tag{7}$$

We observed that web pages that are far away from its genre centroid tend to negatively affect the performance of categorization. Our hypothesis is that these web pages increase web search noise and, consequently, they cannot be considered as useful training pages. For this reason, they should be excluded during centroid computation.
Assume that you have obtained a set of genre centroids $G = \{G_1, \ldots, G_j, \ldots, G_{|G|}\}$, where $|G|$ is the number of genres. In our approach, we discarded web pages that have a similarity with the genre centroid below a predefined threshold s_0.
For each genre g_j, we calculate a new set of training or labeled web pages s_j as follows:

$$s_j = \{p_i \in g_j \backslash sim(p_i, G_j) \geq s_0\} \tag{8}$$

where p_i is a web page and sim is the cosine similarity between the page p_i and the genre centroid G_j defined as follows:

$$sim(p_i, G_j) = \frac{p_i \cdot G_j}{\|p_i\| \cdot \|G_j\|} \tag{9}$$

The sets of training pages obtained after refining will be used to recalculate the genre centroids using the normalized sum presented in equation 7 as follows:

$$S_j = \frac{1}{\|s_j\|} \cdot \sum_{p_i \in s_j} p_i \tag{10}$$

Finally, the refined centroids will be applied to classify new web pages. Note that the complexity of centroid construction is linear to the number of labeled web pages m and to the number of predefined genres $|G|$. Hence, learning time depends on $O(m|G|)$.

In order to choose the appropriate threshold, we carried out the experimental study described in the next subsection.

3.4 Categorization of New Web Pages

In our approach, the categorization of new web pages is performed incrementally. For each new web page p, we calculated its cosine similarity with all genre centroids. Then, we refined the centroids that have a similarity with the page p greater or equal than S_0.

The refining process is performed as follows:

$$NS_i = NS_i + p, S_i = \frac{NS_i}{\|NS_i\|} \tag{11}$$

where NS_i is the non-normalized centroid of the genre g_i and represents the norm of the vector. NS_i is calculated as follows:

$$\|NS_i\| = \sqrt{\sum_{k \in NS_i} wtd_{ijk}^2} \tag{12}$$

The complexity of web page classification is linear to the number of genres $|G|$ and to the number of unlabeled web pages. Therefore, the running time for classification depends on $O(n|G|)$.

3.5 Combination

The basic idea behind the combination of different classifier methods is to create a more accurate classifier via some combination of the outputs of the contributing classifiers. In our approach, the idea is based on the intuition that the combination of homogenous classifiers using heterogeneous features might improve the final result.

OWA operators OWA (Ordered Weighting Average) operators were first introduced in (Yager, 1988). Generally speaking, a mapping $F : [0,1]^n \rightarrow [0,1]$ is called an OWA operator of dimension n if it is associated with a weighting vector $W = [w_1, \ldots, w_i, \ldots, w_n]$, such that $w_i \in [0,1]$, $\sum_i w_i = 1$ and $F(a_1, \ldots, a_n) = \sum_i w_i b_i$, where b_i is the ith largest element in the collection $a_1, \ldots, a_n$. Yager (1988) suggested two methods for identifying weights. The first approach uses learning techniques. The second one firstly gives some semantics to the weights, then based on this semantics, the values for weights are provided.
In the experiments described in this article, we used the second method based on fuzzy linguistic quantifiers for the weights. According to Zadeh (1983), there are two types of quantifiers: absolute and relative. Here, we used relative quantifiers typified by terms such as "as most", "as least half", etc. A relative quantifier Q is defined as a mapping $Q : [0,1] \rightarrow [0,1]$, verifying $Q(0) = 0$, there exists $r \in [0,1]$ such that $Q(r) = 1$ and Q is a non-decreasing function. Herrera and Verdegay (1996) defined a quantifier function as follows:

$$Q(r) = \begin{cases} 0, & if r < a; \\ \frac{r-a}{b-a}, & if r \in [a,b]; \\ 1, & if r > b. \end{cases} \quad (13)$$

where a, $b \in [0,1]$ are two parameters. Yager (1988) computed the weight $w_i (i = 1, \ldots, n)$ as follows:

$$w_i = Q(\frac{1}{n}) - Q(\frac{i-1}{n}) \quad (14)$$

where n is set to 3 because we have three classifiers, named URL, logical and hypertext classifiers. Depending on the values of the parameters a and b, we used the following function operators:

- **Minimum:** Represented by the quantifier "For all" and the function:

$$Q(r) = \begin{cases} 0, & r \neq 1; \\ 1, & r = 1. \end{cases} \quad (15)$$

- **Maximum:** Represented by the quantifier "There exists" and the function:

$$Q(r) = \begin{cases} 0, & r < 1/3; \\ 1, & r \geq 1. \end{cases} \quad (16)$$

- **Median:** Represented by the quantifier "At least one" and the function:

$$Q(r) = \begin{cases} 0, & r < 0; \\ r, & 0 \leq r \leq 1; \\ 1, & r > 1. \end{cases} \quad (17)$$

- **Vote1:** Represented by the quantifier "At least half" and the function:

$$Q(r) = \begin{cases} 0, & r < 0; \\ 2r, & 0 \leq r \leq 0.5; \\ 1, & r > 0.5. \end{cases} \quad (18)$$

- **Vote2:** Represented by the quantifier "As possible" and the function:

$$Q(r) = \begin{cases} 0, & r < 0.5; \\ 2r - 1, & 0.5 \leq r \leq 1; \\ 1, & r > 0.5. \end{cases} \quad (19)$$

Decision templates Decision templates were proposed by Kuncheva et al. (2001).

Let E_1, E_2 and E_3 be the URL, the logical and the hypertext classifiers. Each of these classifiers produces the output $E_i(p) = [d_{i1}(p), \ldots, d_{i|G|}(p)]$ where $d_{ij}(p)$ is the membership degree given by the classifier E_i that a web page p belong to the genre j. The outputs of all classifiers can be represented by a decision profile DP matrix as follows:

$$DP(p) = \begin{pmatrix} d_{11}(p) & \ldots & d_{1|G|}(p) \\ d_{21}(p) & \ldots & d_{2|G|}(p) \\ d_{31}(p) & \ldots & d_{3|G|}(p) \end{pmatrix} \quad (20)$$

Using the training set $Z = Z_1, \ldots, Z_N$, we computed the fuzzy template F of each genre i, which is represented by a $3 \times |G|$ matrix $F_i = f_i(k, s)$. The element $f_i(k, s)$ is calculated as follows:

$$f_i(k,s) = \frac{\sum_{j=1}^{N} Ind(Z_j,i) \cdot d_{ks}(Z_j)}{\sum_{j=1}^{N} Ind(Z_j,i)} \quad (21)$$

where $Ind(Z_j,i)$ is an indicator function with value 1 if Z_j comes from genre i and 0 otherwise. At this stage, the ranking of genres can be achieved by aggregating the columns of DP using fixed rules (minimum, maximum, product, average, etc.). Another method calculates a soft class label vector with components expressing similarity S between the decision template DP and the fuzzy template F. The final classification CLV is defined as follows:

$$CLV(p) = [\mu_1(p), \ldots, \mu_i(p), \ldots, \mu_{|G|}(p)] \quad (22)$$

where $\mu_i(p)$ is the similarity $S(F_i, DP(p))$ between the fuzzy template F_i of the genre i and the decision profile $DP(p)$ of the web page p. This similarity is calculated using the Euclidean measure as follows:

$$\mu_i(p) = S(F_i, DP(p)) = 1 - \frac{1}{3 \times |G|} \cdot \sum_{k=1}^{3} \sum_{s=1}^{|G|} (f_i(k,s) - d_{ks}(p))^2 \quad (23)$$

4 Corpora

In our experiment, we used the KI-04 corpus and WebKB collection[5]. These corpora are composed of English web pages. Each web page is associated with a specific source URL address, and belongs to a single genre class.

- **KI-04** corpus was compiled by Meyer Zu Eissen and Stein (Meyer Zu Eissen and Stein, 2004). It is composed of 1205 HTML web pages, which are divided into eight genres (see Table 1).

- **WebKB** corpus was created at Carnegie-Mellon University during the WebKB project (Craven et al., 1998). This corpus contains 4249 HTML web pages from four different universities. The corpus comprises six genres (see Table 2).

[5]Both these corpora can be reached through the WebGenreWiki `http://http://www.webgenrewiki.org/index.php5/Genre_Collection_Repository/`

Table 1: Composition of the KI-04 corpus

Genre	♯ of web pages
Article	127
Download	151
Link collection	205
Private portrayal	126
Non-private portrayal	163
Discussion	127
FAQ	139
Shop	167

Table 2: Composition of the WebKB corpus

Genre	♯ of web pages
Student	1541
Faculty	1063
Staff	126
Department	170
Project	474
Course	875

5 Evaluation

In this section, we describe our evaluation within the $FRICC$ framework. $FRICC$ is the abbreviation of Flexible, Refined and Incremental Centroid-based Classifier. The aims of the evaluation process can be summarized as follows:

- Identify the best proportions of labeled and unlabeled web pages to achieve the best performance.
- Identify the appropriate number of terms to obtain the best performance.
- Identify the appropriate values of normalization parameters.
- Identify the best thresholds.
- Identify the best combination techniques.

For multiclass corpora, it is suitable to use the break-even-point (BEP), which is defined in terms of the standard measures of precision and recall (Joachims, 1997). Precision P is the proportion of true document-category assignments among all assignments predicted by the classifier. Recall R is the proportion of true document-category assignments that were also predicted by the classifier. Formally, the BEP statistic finds the point where precision and recall are equal. Since this is hard to achieve in

practice, a common approach is to use the arithmetic mean of recall and precision as an approximation, i.e. $BEP = (P + R)/2$. Since our corpora are unbalanced, we used the micro-averaged BEP computed by first summing the elements of all binary contingency tables (one for each genre). Then, the micro-averaged BEP is computed from these accumulated statistics.
Note that the noisy web pages are not considered in evaluation process.

To measure the performance, we used the $10 \times k$ cross-validation. This means that we randomly split each corpus into k equal parts. Then we used one part for testing and the remaining parts for training. This process was performed 10 times and the final performance is the average of the 10 individual performances. The number k is identified experimentally according to the used features and corpora.

5.1 Results

In the following paragraphs, we describe a number of experiments and show the results.

Effect of Incremental Aspect In this experiment, we varied the proportion of unlabeled web pages between 10% and 90% by step of 10%. For each proportion, we measured the micro-averaged BEP for each feature set and corpus. The results are illustrated in Figure 1.

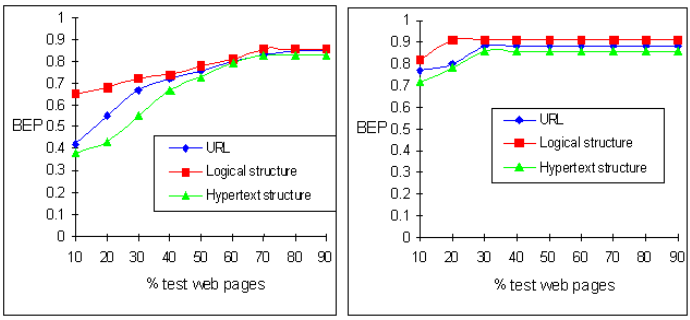


Figure 1: Micro-averaged BEP for each feature and for both KI-04 (Left) and WebKB (Right) corpora when the proportion of test pages is varied between 10% and 90%

The curves presented in Figure 1 shows that micro-averaged BEP depends on the proportion of labeled and unlabeled web pages. These curves also show that it is the logical structure classifier that achieves the best performance for both KI-04 and WebKB corpora. The proportions of unlabeled and labeled web pages to achieve the best performance are presented in the Table 3. These proportions are used in the next experiments.

Table 3: Best proportions of training and test web pages (Test%-Train%)

	URL	Logical Structure	Hypertext Structure
KI-04	80%-20%	70%-30%	70%-30%
WebKB	30%-70%	20%-80%	30%-70%

Effect of Vocabulary Size The aim of this experiment is to identify the ideal number of terms to achieve the best performance. For this purpose, we calculated the micro-averaged BEP by varying the number of terms between 5 and 3000. The number of terms complies to the information gain measure. Note that in this experiment, we used the $tfidf$ weighting technique without normalization. The obtained results are illustrated in Figure 2.

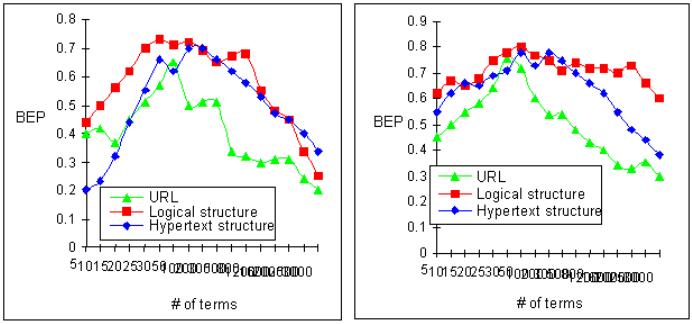


Figure 2: Micro-averaged BEP for each feature and for both KI-04 (Left) and WebKB (Right) corpora when the number of terms is varied between 5 and 3000

The ideal number of terms to achieve the best performance are summarized in Table 4. These values will be used in the next experiment.

Table 4: Best values of number of terms

	URL	Logical Structure	Hypertext Structure
KI-04	100	50	200
WebKB	50	100	300

Effect of Term Weighting In order to evaluate the effect of each normalization factor alone ($icsd$, csd and sd), we conducted an experiment whose results are showed in Table 5.

Table 5: The effect of each normalization factor on genre categorization performance

KI-04					
α	β	γ	**URL**	**Logical**	**Hypertext**
1	0	0	0.63	0.74	0.66
-1	0	0	0.66	0.75	0.68
0	1	0	**0.68**	**0.76**	0.72
0	-1	0	0.67	0.68	0.68
0	0	1	0.64	0.63	0.7
0	0	-1	0.66	0.68	**0.73**
WebKB					
1	0	0	**0.78**	**0.83**	**0.81**
-1	0	0	0.75	0.81	0.76
0	1	0	0.72	0.78	0.70
0	-1	0	0.70	0.75	0.65
0	0	1	0.65	0.80	0.64
0	0	-1	0.71	0.72	0.58

We observed that the *icsd* factor is very suitable for the KI-04 corpus because it contains heterogeneous genres. On the other hand, the *sd* factor achieves the best performance for the WebKB corpus because it contains homogenous genres.

Table 6: The effect of normalization on genre categorization performance

KI-04					
α	β	γ	**URL**	**Logical**	**Hypertext**
0.5	1	0.5	0.66	0.72	0.7
-0.5	0.5	1	0.45	0.78	0.75
0.5	-1	0	0.72	0.8	0.67
0.5	-0.5	-0.5	0.6	0.81	0.63
0.5	0	-0.5	0.68	0.73	0.55
-1	0.5	-0.5	0.7	0.65	0.77
0.5	**-1**	**-0.5**	**0.7**	**0.81**	**0.82**
1	0	-0.5	0.66	0.55	0.8
-1	-0.5	1	0.7	0.52	0.81
WebKB					
0.5	1	0.5	0.76	0.68	0.43
-0.5	0.5	1	0.77	0.66	0.71
0.5	-1	0	0.83	0.73	0.74
0.5	-0.5	-0.5	0.85	0.45	0.55
0.5	0	-0.5	0.81	0.65	0.45
-1	0.5	-0.5	0.56	0.76	0.82
0.5	**-1**	**-0.5**	**0.86**	**0.86**	**0.84**
1	0	-0.5	0.83	0.77	0.68
-1	-0.5	1	0.65	0.58	0.52

To choose the appropriate value of normalization parameters, we varied the values of α, β and γ between -1 and 1 by a step of 0.5. The best results are presented in the

Table 6. The best performance is reported by setting the normalization parameters α, β and γ to 0.5, -1 and -0.5 respectively. These values will be used in the next experiment to choose the appropriate threshold.

Effect of Refining Aspects To measure the effect of refining on genre categorization, we varied the refining threshold between 0 and 1 by step of 0.1. Zero value means that is no refining. As illustrated in Figure 3, the value of threshold affects the micro-averaged BEP of genre categorization.

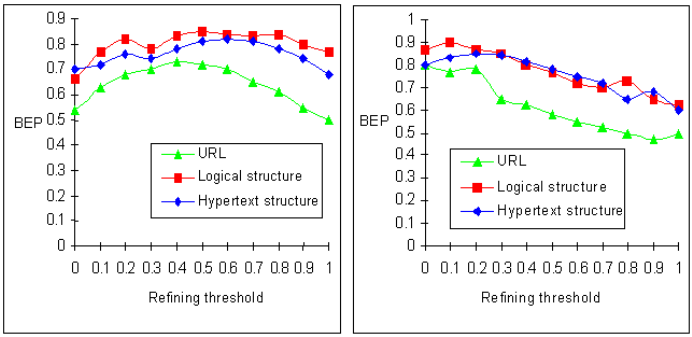


Figure 3: Micro-averaged BEP for each feature and for both KI-04 (Left) and WebKB (Right) corpora when the refining threshold is varied between 0 and 1

We noticed that in the case of noisy web pages like those contained in KI-04 corpus, the refining is very useful. On the other hand, for noiseless corpus like WebKB corpus, the refining is useless. The best refining thresholds will be used in the next experiments. The number of noisy web pages and the thresholds to achieve the best micro-averaged BEP are shown in Table 7.

Table 7: Best number of noisy web pages and refining thresholds for each feature and corpus

	URL	Logical Structure	Hypertext Structure
KI-04	7(0.4)	6(0.5)	4(0.6)
WebKB	5(0)	4(0.1)	7(0.2)

Effect of Combination Here we conducted many experiments to choose the appropriate operator for combination. The obtained results are shown in the Table 8. These results show that the decision template technique provides the best micro-averaged BEP (0.96 for KI-04 corpus and 0.98 for WebKB corpus).

Table 8: Micro-averaged BEP for each combination technique and for both KI-04 and WebKB corpora

Combination technique	KI-04	WebKB
Minimum	0.88	0.93
Maximum	0.96	0.97
Median	0.91	0.94
Vote1	0.90	0.94
Vote2	0.90	0.93
Decision templates	**0.96**	**0.98**

6 Comparisons

Accuracy The majority of the previous studies do not provide a reliable comparison with other approaches. The main reason for this is that, until recently, there were no publicly available and standard corpora for this task. Another reason is that there is not a commonly perceived sense of specific web page genres. For example, in two recent studies, user agreement was only 57In this article, we propose a comparison with other experiments, namely Meyer Zu Eissen and Stein (2004), Kanaris and Stamatatos (2007) and Santini (2007), where the KI-04 corpus is employed. The WebKB corpus is used only by Boese and Howe (2005), so we will compare our results with this experiment.

These experiments are evaluated using the accuracy measure. The micro-averaged accuracy for both KI-04 and WebKB corpora for each author is presented in Table 9. According to the results shown in this table, our approach outperforms other methods.

Table 9: Micro-averaged accuracy for both KI-04 and WebKB corpora

Author	KI-04	WebKB
Meyer Zu Eissen and Stein (800 web pages)	0.70	-
Boese and Howe	0.75	0.80
Kanaris and Stamatatos (1205 web pages)	0.84	-
Santini (1205 web pages)	0.70	-
Our approach	**0.96**	**0.98**

Machine Learning Techniques Since our approach is based on new learning aspects, we conducted experiments to compare it against other known machine learning methods used in genre categorization. Among these techniques, we used the Rocchio algorithm (Rocchio with α=14 and β=4 as control parameters), K-Nearest Neighbor (KNN with k=5), Support Vector Machines (SVM with Fisher Kernel), Naïve Bayes (NB) and decision trees (TreeNode). These techniques are implemented within the Rainbow toolkit. Micro-averaged BEP for each feature set and for both KI-04 and WebKB corpora are presented in Tables 10 and 11.

Table 10: Micro-averaged BEP for the KI-04 corpus

	URL	Logical structure	Hypertext structure
FRICC	**81.12**	**85.44**	83.17
SVM	80.76	84.75	**84.20**
Rocchio	77.55	82.10	82.15
NB	71.95	79.65	81.45
KNN	67.35	65.85	80.56
TreeNode	62.77	61.89	65.55

Table 11: Micro-averaged BEP for the WebKB corpus

	URL	Logical structure	Hypertext structure
FRICC	**85.73**	**87.32**	84.24
SVM	84.33	86.88	82.29
Rocchio	82.18	87.24	**88.78**
NB	80.88	86.76	80.85
KNN	70.22	74.11	76.16
TreeNode	61.89	64.40	62.33

Statistical Significance To determine the statistical significance of the results, we used 5×2 cross validation $t-test$ (Dietterich, 1998). The results are presented in Table 12. The symbols used in this table are defined as follows:

- $\approx$ Indicates no significant differences.
- $<$ Indicates that the machine learning method achieves a significantly lower measurement than $FRICC$ with 0.05 as a significance level.
- $<<$ Indicates that the machine learning method achieves a significantly lower measurement than $FRICC$ with 0.01 as a significance level.
- $<<<$ Indicates that the machine learning method achieves a significantly lower measurement than $FRICC$ with 0.005 as a significance level.

Table 12 shows that the $FRICC$ approach outperforms all other machine learning methods in 27 cases. Only SVM has similar performance to $FRICC$.

Training and Test Times Here we consider another important aspect, namely execution speed. Time is a very important aspect, especially when genre classification has to be integrated in a search engine. Figures 4, 5 and 6 show a comparison of the execution speeds for each classification method, in both training and test phases, for the KI-04 and WebKB corpora.

Results show that our approach is the fastest, but also Rocchio and SVM have a good performance. These results indicate that the required time is proportional to the number of categories instead of the number of web pages. Decision tree is, indeed, the slowest machine learning technique for all feature sets and for both corpora.

Table 12: Statistical Significance of our approach $FRICC$ against other machine learning techniques

KI-04			
	URL	**Logical Structure**	**Hypertext Structure**
SVM	$\approx$	$<<$	$<<$
Rocchio	$<<$	$<$	$<<$
NB	$<<$	$<<$	$<<<$
KNN	$<<$	$<<$	$<<<$
TreeNode	$<<<$	$<<<$	$<<<$
WebKB			
SVM	$\approx$	$<<$	$\approx$
Rocchio	$<<$	$<$	$<<$
NB	$<<$	$<<$	$<<<$
KNN	$<<<$	$<<$	$<<<$
TreeNode	$<<<$	$<<<$	$<<<$

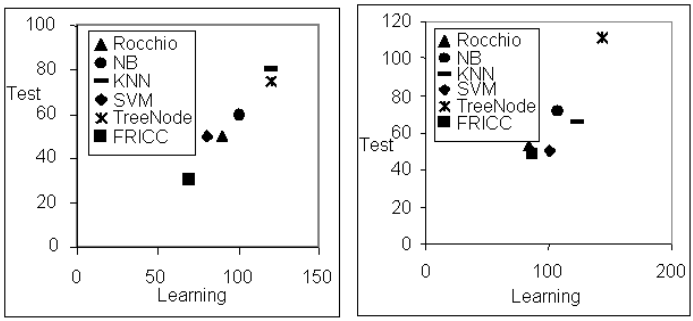


Figure 4: Training and test times for URL and for both KI-04 (left) and WebKB (right) corpora

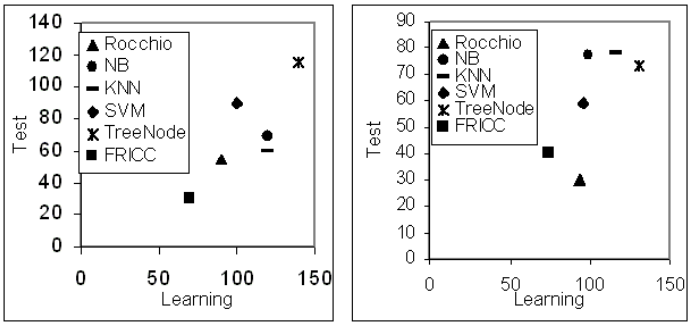


Figure 5: Training and test times for logical structure and for both KI-04 (left) and WebKB (right) corpora

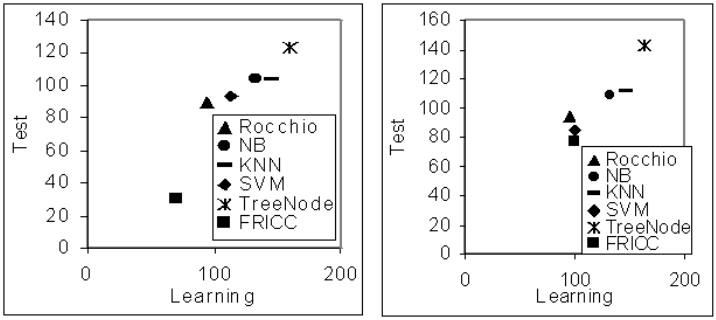


Figure 6: Train and test times for hypertext structure and for both KI-04 (left) and WebKB (right) corpora

7 Conclusions and Future work

In this article, we proposed a new approach for genre categorization of web pages. Our approach implements four new aspects that were not explored in previous studies on genre categorization. These aspects are flexibility, refining, incrementing and combination. Additionally, we conducted many experiments to measure the effectiveness, efficiency and speed of these aspects. Comparisons with previous approaches shows that our method is very fast and outperforms results documented in previous work.

In the future we hope to investigate the following points:

- As the pdf format is a very useful format on the web, we propose to classify pdf documents.
- In this work, we used only English web page, in the future we wish to focus on Arabic web documents.
- As our approach is very fast and outperforms many other machine learning techniques, we hope to include it in a search engine (i.e. Google, FireFox), in a similar way as the WEGA add-on (Stein et al.) .

Remark

The work described in this article summarizes the PhD thesis "Catégorisation Flexible et Incrémentale avec Raffinage de Pages web par Genre", completed by the author, Chaker Jebari, in October 2008, *Tunis El Manar University*, College of Science, Computer Science Department, Tunisia.

Acknowledgements

The author would like to thank the anonymous reviewers, the proofreaders and the journal's editors for useful comments, for the stylistic improvement of the article, and for the considerable editorial effort invested in the publication of this work.

References

Argamon, S., Koppel, M., and Avneri, G. (1998). Routing documents according to style. In *Proc. International Workshop on Innovative Internet Information Systems (IIIS-98)*, Pisa.

Boese, E. S. (2005). *Stereotyping the Web: Genre Classification of Web Documents (M.S. Thesis)*. Computer Science Department, Colorado State University, USA.

Boese, E. S. and Howe, A. E. (2005). Effect of web document evolution on genre classification. In *Proceedings of the 14th ACM International conference on Information and knowledge management*.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. (1998). Learning to extract symbolic knowledge from the word wide web. In *Proceedings of the 10th conference on artificial Intelligence*.

Dewdney, N., VanEss-Dykema, C., and MacMillan, R. (2001). The form is the substance: Classification of genres in text. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*, 10(7):1895–1923.

Finn, A. (2002). *Machine learning for genre classification (M.S. Thesis)*. Computer Science Department, University College of Dublin, UK.

Finn, A. and Kushmerick, N. (2003). Learning to classify documents according to genre. In *Proceedings of the Workshop "DOING IT WITH STYLE: Computational Approaches to Style Analysis and Synthesis*, Mexico.

Herrera, F. and Verdegay, J. L. (1996). *Genetic algorithms and soft computing*. PhysicaVerlag, Heidelberg, Germany.

Joachims, T. (1997). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of ICML-97*, pages 143–151.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of 10th European Conference on Machine Learning*.

Kanaris, I. and Stamatatos, E. (2007). Webpage genre identification using variable length character n-grams. In *Proceeding of the 19th IEEE International Conference on Tools with Artificial Intelligence*.

Karlgren, J. (1999). Stylistic experiments in information retrieval. In *Natural Language Information Retrieval*.

Kennedy, A. and Shepherd, M. (2005). Automatic identification of home pages on the web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th ACL/8th EACL*, pages 32–38.

Kuncheva, L. I., Bezdek, J. C., and Duin, R. P. (2001). Decision templates for multiple classifier fusion. *Pattern Recognition*, 34(2):299–314.

Lertnattee, V. and Theeramunkong, T. (2004). Effect of term distributions on centroid-based text categorization. *Journal of Information Sciences*, 158(1):89–115.

Lim, C. S., Lee, K. J., and Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Journal of Information processing and management*, 41(5):1263–1276.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Mehler, A., Geibel, P., and Pustylnikov, O. (2007). Structural classifiers of text types: Towards a novel model of text representation. *LDV Forum*, 22(2):51–65.

Mehler, A., Sharoff, S., and Santini, M., editors (2009). *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Meyer Zu Eissen, S. (2007). *On Information Need and Categorizing Search (PhD. Thesis)*. University of Paderborn.

Meyer Zu Eissen, S. and Stein, B. (2004). Genre classification of web pages: User study and feasibility analysis. In *Proceedings KI 2004: Advances in Artificial Intelligence*, pages 256–269.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rauber, A. and Müller-Kögler, A. (2001). Integrating automatic genre analysis into digital libraries. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10.

Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Santini, M. (2007). *Automatic identification of genre in web pages (PhD. Thesis)*. University of Brighton, UK.

Shepherd, M., Watters, C., and Kennedy, A. (2004). Cybergenre: automatic identification of home pages on the web. *Journal of Web Engineering*, 3(3):236–251.

Shepherd, M. A. and Watters, C. (1998). Evolution of cybergenre. In *Proceedings of the 31nd Hawaiian International Conference on System Sciences*.

Stamatatos, E., Fokatakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics*.

Stein, B., Meyer zu Eissen, S., and Lipka, N. Web genre analysis: Use cases, retrieval models, and implementation issues.

Vapnik, V. (1995). *The Nature of Statistical Learning*. Springer-Verlag.

Yager, R. (1988). On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97*.

Zadeh, L. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems*, 11(3):199–227.

Vedrana Vidulin, Mitja Luštrek, Matjaž Gams

Multi-Label Approaches to Web Genre Identification

A web page is a complex document which can share conventions of several genres, or contain several parts, each belonging to a different genre. To properly address the genre interplay, a recent proposal in automatic web genre identification is multi-label classification. The dominant approach to such classification is to transform one multi-label machine learning problem into several sub-problems of learning binary single-label classifiers, one for each genre. In this paper we explore multi-class transformation, where each combination of genres is labeled with a single distinct label. This approach is then compared to the binary approach to determine which one better captures the multi-label aspect of web genres. Experimental results show that both of the approaches failed to properly address multi-genre web pages. Obtained differences were a result of the variations in the recognition of one-genre web pages.

1 Introduction

A web page is a complex document which can share conventions of several genres or contain several parts each of different genre. While this is recognized in the community of automatic web genre identification (AWGI), state-of-the-art implementations of genre classifiers mostly use single-label classification scheme (Karlgren and Cutting, 1994; Lim et al., 2005). In other words, they attribute to a web page one genre label from the set of predefined labels. Recent line of research (Santini, 2007, 2008), however, showed that multi-label classification scheme is more suitable for capturing the web page complexity. In our study we follow this scheme with some modifications.

The need for attributing more that one genre label to a web page is noticed by several authors (Roussinov et al., 2001; Meyer zu Eissen and Stein, 2004; Rosso, 2005), however, the primary goal of this studies was not the implementation of multi-label genre model. In contrast, Santini (2007) implemented the model based on zero-to-multi genre assignment. Her classification scheme was motivated by the two characteristics of genre: hybridism and individualization (Santini, 2008). Since several genres are easily combined in a single web page, she argues that such hybrid forms require attribution of multiple genre labels. In contrast, the absence of the mechanisms which would force the strict application of genre conventions, as in the case of e.g. scientific article, allows the individualization of genres. The individualized web pages have unclear genre conventions, and are consequently marked with zero genres. The model based on the presented scheme was implemented in two steps. First, the combinations of

facets[1] representing text types were hard-coded to obtain a middle-layer model for the recognition of text types. Text types were inferred using the modified form of Bayes' theorem, namely the odds-likelihood or subjective Bayesian method. Second, if-then rules were created to identify genres from the combination of text types and other features (e.g. linguistic, HTML).

Another model of Stubbe et al. (2007) was partially concerned with the issue of multi-label approach to AWGI. They built multiple genre-specific classifiers, one per genre, by combining features provided by the genre expert into rules. They pointed that the classifiers can be combined into a scheme which can attribute several genre labels to a web page. An improved classifier's precision was observed. However, they did not explore this issue in more detail since their main task was to exploit interdependencies between genre specific classifiers to improve the precision of a single label assignment.

In contrast to the presented approaches, which are based on expert knowledge, our goal is to induce a multi-label model from the example web pages with supervised machine learning (ML) methods. Learning a multi-label model can be achieved through the problem transformation or through the algorithm adaptation approach (Tsoumakas and Katakis, 2007). We follow problem transformation approach, and explore two transformations: a transformation to a multi-class problem and a transformation to a set of binary sub-problems. Finally, we use standard ML algorithms to learn the models from the transformed data, and we test their performances to understand which type of the model can better deal with multi-genre web pages.

The binary approach shares several characteristics with the approaches of Santini (2007) and Stubbe et al. (2007). First, we introduce a set of binary classifiers, one per genre. Second, by combining positive answers of multiple classifiers we realize zero-to-multi genre assignment, attributing zero genres when there are no positive answers, a single genre when there is only one positive answer, and multiple genres when there are several positive answers. In contrast to the related work, we do not analyze the performance of separate binary classifiers. Instead, we focus on evaluating the performance of multi-label classifier as a whole.

Our multi-class approach differs from previous research. Its main advantage is the induction of a single classifier, for which we assume that it can better learn the overlaps between different genres, and consequently better recognize web pages containing multiple genres. There are certain disadvantages to this approach. One lies in the inability to capture all genre combinations within a single corpus. The classifier, therefore, cannot properly recognize a web page containing a new genre combination. The best result which the classifier can produce in such a situation is to recognize the dominant genre. Another problem lies in the inability to properly define attribution of zero labels. Even if we introduce the label, e.g. "N/A", the question is what kind

[1] Santini (2007) defines facet as "an 'aspect' in the communicative context that is reflected in the use of language". For example, first person facet is complex feature accounting for appearance of first person pronouns in a web page, and indicates the communication context related to the text producer.

of example web pages can we put in this category to properly learn it. This problems should be explored in more detail, which is beyond the scope of this paper.

The two approaches were tested on the multi-label 20-Genre-Collection corpus, collected for the purpose of learning the classifier for implementation into a search engine. Since there is no common, widely agreed upon set of web genre categories (Rehm et al., 2008), for that purpose we defined 20 broad categories which combined into the multi-label scheme tend to be robust enough to deal with the diversity and the complexity of web pages on the open web. The corpus is composed of web pages in English, therefore examples in ML tasks are individual web pages.

The rest of the paper is organized as follows. In Section 2 we describe the web genre categories and in Section 3 the multi-label corpus we experimented with. Section 4 lists the features used to describe web pages in terms of web genre. Section 5 presents the methodology behind experiments. Section 6 presents experimental results with discussion, and Section 7 concludes the paper and presents the directions for future work.

2 Web Genre Categories

Although genres form taxonomy, for practical purposes this taxonomy is usually reduced to one level. Santini (2007) selected only basic level genres that can be directly instantiated by formulating text in the proposed genre, e.g. *Personal home page* or *FAQ*. Lim et al. (2005) used broad categories of higher level, composed of one or more basic level genres. For example, the genre *Journalistic materials* includes press reportage, editorial and review, while the genre *Informative materials* includes recipes, lecture notes and encyclopedic information. The advantage of the second approach and the reason while we are following it is that it seems more natural to cope with the diversity of the Internet. However, the disadvantage lies in the difficulty to represent common characteristics of web pages that compose such broad categories. Therefore, the genre classifiers learned on corpora with broad categories showed somewhat lower performance.

We defined our set by reusing and refining existing sets and by adding new categories. The starting point was the work of Lim et al. (2005). In total, Lim et al. (2005) selected 16 categories, 8 non-textual (*Personal homepages, Public homepages, Commercial homepages, Bulletin collections, Link collections, Image collections, Simple table/lists, Input pages*) and 8 textual (*Journalistic materials, Research reports, Official materials, Informative materials, FAQs, Discussions, Product Specifications, Others (informal texts)*). We defined 20 categories presented in the first column of Table 1. The overlaps with the Lim et al.'s (2005) categories are presented in the last column. Partial overlap is marked with an asterisk (*).

Childrens' category includes multiple genres aimed at younger audience. Common characteristics of pages belonging to this category are the use of simple language and colorful formatting. The identification of this category could be useful for children and probably even more for their parents. *Commercial/promotional* pages have the common purpose of promoting organizations and selling one's own products and services. In

contrast, selling products of others is the purpose of *Shopping* pages. *Community* page has the purpose of involving a visitor in the creation of the page, usually by contributing content in a limited way (forums), although there are pages where the users are given even more freedom. *Content delivery* page delivers content that is not part of the page (e.g. download pages). Another purpose is to present embedded non-textual content (e.g. page with flash game). Lim et al.'s (2005) *Image collections* category has a similar purpose. The difference is that we broaden it by taking into consideration other types of media. Furthermore, it is difficult to include genres such as jokes and horoscopes in commonly used genre categories. Their common purpose, which could be interesting to a search engine users, is to entertain. Therefore, we added *Entertainment* category. *Error message* pages are not particularly interesting to a visitor, and are not intended to be offered as a choice in a search engine. Instead, this category is used to filter such uninteresting pages. The purpose of *Gateway* is to transfer the visitor to another page. Some of the gateway pages (e.g. login page) are overlapping with Lim et al.'s (2005) *Input pages*. *Official pages* partially overlap with Lim et al.'s (2005) *Official materials*. For example, they labeled legal info and copyright materials pages as *Official materials*, and we would label them as *Official*. In contrast, they labeled ad page as *Official materials*, while we would label it as *Shopping*. *Personal* pages are home-made (not professionally formatted), written in informal and subjective manner. This category includes Lim et al.'s (2005) *Personal homepages* and opinions, which Lim et al. (2005) classified as *Discussions*. *Poetry* and *Prose fiction* are genres considered by Lim et al. (2005) as informal texts. Because a search engine user can have special interest in those genres (e.g. searching for lyrics of Madonna's song) we separated them in two distinct categories. Similar to *Childrens'* pages, *Pornographic* covers multiple genres. It is, however, targeted at the adult audience.

3 20-Genre Collection Corpus

We were not able to obtain a publicly available corpus, which adequately represents genre categories presented in Table 1. Therefore, we built our own corpus.

The web pages were collected from the Internet using three methods. Firstly, we used highly-ranked Google hits for popular keywords (e.g. "Britney Spears"). The keywords were chosen according to 2004 Year-End Google Zeitgeist statistics (`http://www.google.com/press/zeitgeist2004.html`). Our purpose was to build a classifier that will not have a problem with recognizing the most popular web pages, which people actually search for. 150 pages were collected by entering the most popular queries from each of the five categories from 2004 Google Zeitgeist. The collected pages ranked from 31^{st} to 60^{th} place. The first 30 hits were skipped as suggested by Lim et al. (2005) to avoid too many *Commercial/promotional* pages. In order to increase the diversity while retaining the popularity criterion, we input the most popular queries of each weekly 2004 Google Zeitgeist into Google and used the hits ranked from 31^{st} to 35^{th} (or 31^{st} to 40^{th} for pages topping the weekly Zeitgeist twice). This gave us 245 pages for a total of 395 pages. Secondly, we gathered 300 random web pages using Mangle

(http://www.mangle.ca/), a random link generator. Finally, we specifically searched for web pages belonging to the genres under-represented to that point by inputting genre-related queries into Google and using relevant hits. The purpose of the last step was to obtain a balanced corpus that represents all genres equally well. Imbalance usually causes difficulties in learning the under-represented classes. In total, 1,539 web pages in English were collected.

The corpus was manually labeled with genres by two independent annotators. Their labels disagreed on about a third of the web pages in the corpus, so those were reassessed by a third and sometimes even a fourth annotator.

The distribution of web pages over the 20 categories in the 20-Genre Collection corpus (http://dis.ijs.si/mitjal/genre/) is presented in Table 2 (the multi-label aspects of corpus are discussed in the section on learning a multi-label genre classifier). The targeted average was 100 pages per genre. We ended up with at least 55 pages of each genre and around 200 pages belonging to the most common genres. Such differences can be attributed to search engine's bias towards certain genres (e.g. *Index, Informative, Journalistic*), or because some genres are simply more common on the Internet.

4 Features

We selected a broad set of features from previous studies and combined them with features obtained from the analysis of 20-Genre Collection corpus to cover different aspects of genre: content, linguistic and visual form, and the context of a web page. In total, 2,491 features were chosen separated in four groups: surface, structural, presentation and context features.

4.1 Surface Features

Surface features pertain to content of a web page. For example, frequent appearance of function word "you" can characterize promotional pages of *Commercial/promotional* genre. They are easily extractable and, hence, commonly used (Stamatatos et al., 2000; Dewdney et al., 2001; Lim et al., 2005). This group includes function words, genre-specific words, punctuation marks, classes of words (such as dates, times, postal addresses and telephone numbers), and word, sentence and document length.

We selected 411 surface features, presented in Table 3. The set of 321 genre-specific words was obtained by combining the list of most frequent content words from the corpus with manually selected genre-describing words. They were stemmed by the Porter stemming algorithm (Porter, 1980).

4.2 Structural Features

Structural features describe syntactic choices. For example, high frequency of nouns can indicate *Informative* pages. They include features like parts of speech (POS), phrases (e.g. noun phrase or verb phrase) and sentence types (e.g. the frequencies of declarative, imperative and question sentences) (Santini, 2007).

We selected 1,908 structural features, presented in Table 4. POS tags were extracted with TreeTagger (Schmid, 1994). Beside single POS, we also extracted POS trigrams to capture pieces of syntactic constructions. To obtain the set of discriminative POS trigrams, we discarded too common and too rare trigrams (Santini, 2004) in two steps. First, we extracted only trigrams that are present more than three times in a web page. Second, we discarded 25% of the most frequent and 25% of the least frequent trigrams in the corpus.

4.3 Presentation Features

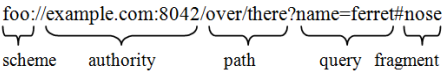
Presentation features describe the formatting of a document. For example, appearances of tags <form> and <input> can indicate *User Input* pages. This group includes token type (e.g. the percentage of a document taken by numbers or whitespaces), text formatting (e.g. amount of bolded text), graphical elements (e.g. the frequencies of images or tables) and similar (Lim et al., 2005).

We selected 93 presentation features, presented in Table 5. Token type can be used to describe formatting of any textual document, while HTML features are specific to web pages. Both single HTML tags and the groups of tags were considered. Following an idea to group tags in macro-features presented in (Santini, 2007), we grouped tags into five categories according to their functionalities.

4.4 Context Features

Context features describe the context in which a web page was found. Under context we assume URL and hyperlinks contained within the web page. URL features describe the structure and the content of URL, while hyperlink features describe types of hyperlinks. For example, appearances of words "blog" and "archive" in URL can indicate *Blog* page, while high number of hyperlinks to a different domain can indicate *Commercial/promotional* pages.

We selected 79 features, 76 URL (Table 6) and 3 hyperlink features (Table 7). The choice of URL features describing its structure follows URL syntax defined by Berners-Lee et al. (1998):



URL content was analyzed by marking the appearances of 54 words most commonly present in URL. The words were stemmed with the Porter stemming algorithm.

5 Learning a Multi-Label Genre Classifier

5.1 Data Set

The data set, to which we will refer to as 20-Genre-Collection data set, was obtained by extracting 2,491 features from 1,539 web pages in the 20-Genre-Collection corpus. All features except those pertaining to URL were expressed as ratios. Since it is more probable that a certain feature would appear more frequently in longer pages, expressing features as ratios eliminates the influence of page length.

From 1,539 web pages, 1,059 are labeled with one, 438 with two, 39 with three and 3 with four labels. On average, there are 1.34 labels per web page.

5.2 Problem Transformations

Multi-label classification assumes association of examples with a set of labels $Y \subseteq L$, L representing the set of labels present in a data set. There are two approaches to multi-label classification: problem transformation and algorithm adaptation (Tsoumakas and Katakis, 2007). We have chosen problem transformation because it allows use of the existing tools for single-label classification. Two transformations are explored in this paper: a transformation to a multi-class problem and a transformation to a set of binary problems.

The multi-class transformation assumes treatment of different sets of labels as distinct single labels. Therefore, the goal is to learn a classifier $F : X \rightarrow P(L)$, where X represents examples and $P(L)$ the power set of L. When applied to the 20-Genre-Collection data set, the categories as *Blog, Childrens', Childrens'-Informative, Community-Informative* were obtained. This transformation explicitly captures overlaps between genres, with the negative side-effect of producing high number of categories. In some cases, newly obtained categories were represented with only one example. To properly train and test a classifier, we removed all the examples labeled with the categories not represented with at least one example in the both train and test sets.

The binary transformation assumes learning $|L|$ binary sub-classifiers $F_l : X \rightarrow \{l, \neg l\}$, one for each label $l \in L$. For example, the 20-Genre-Collection data set is transformed into 20 data sets each containing all the examples of the original data set, labeled as positive (e.g. "1") if the labels of the original example contained l and as negative (e.g. "0") otherwise.

5.3 Learning Classifiers

On the transformed data we applied LibSVM (Fan et al., 2005) and AdaBoost (Freund and Schapire, 1996) to learn the classifiers.

LibSVM has built-in problem transformation functionalities, and is good at handling high number of features and sparse data. In the process of tuning the algorithm, we followed the recommendations of Hsu et al. (2008). First recommendation is to scale the data to avoid features in greater numeric ranges to dominate those in smaller

numeric ranges. We scaled the feature values to fall into the [0, 1] interval. Second recommendation is to test the RBF kernel $\left(K\left(x_i, x_j\right) = exp(-\gamma||x_i - x_j||^2), \gamma > 0\right)$ first since it can handle the cases where the relation between class labels and attributes is nonlinear. Besides, they argue that the linear kernel is a special case of RBF kernel, which is a logical second step to test since it is good at handling the problems with higher number of features in comparison to the number of examples. We compared the performances of the two kernels on our data, and the use of linear kernel instead of the RBF did not result in any improvement at all. Third recommendation is to select the parameters C and γ with the grid search in the space of models induced with exponentially growing sequences of parameters C and γ. We used the tool contained within the LibSVM, and evaluated the quality of parameters using the 3-fold cross-validation on the training set. The parameters of a model with the best cross-validation accuracy were picked. In the case of binary transformation the choice of the parameters was separately done for each sub-classifier.

AdaBoost is a meta-learning algorithm. In our previous research (Vidulin et al., 2007) we boosted J48 decision trees (Witten and Frank, 2005), and obtained the best performance among five algorithms (sequential minimal optimization, Naïve Bayes, J48 decision trees, Random Forrest and AdaBoost) tested on the binary transformation of the 20-Genre-Collection data set.

5.4 Evaluation

The performance of multi-label classifiers was evaluated using stratified 3-fold cross--validation. Stratification is a problem which can be approached in different manners in multi-label setting. One approach is to do the problem transformations first and than to separate the data into folds. It results in better balance of classes on the level of individual ML sub-problems. The second approach is to separate the folds before problem transformations and to stratify in a manner to obtain equal distribution of single genre categories over the folds. Since our goal was to obtain the same train-test splits to allow comparisons between induced classifiers we used the second approach. Considering that the number of examples per category decreases after multi-class transformation, we used three instead of ten folds to increase the chance of obtaining more test examples per class.

The performance of classifiers was evaluated with several measures: exact match ratio, micro-averaged precision, recall and F-measure, and macro-averaged precision, recall and F-measure.

Exact match ratio (EX) counts exact matches between the predicted and the actual labels (Eq. 1). This measure is, in a way, similar to accuracy in the case of the single-label classification. However, it does not account for e.g. two out of three correctly predicted labels which is fairly good success in the multi-label setup.

$$EX = \frac{\sum_{i=1}^{M} I[Y_i^{\text{predicted}} = Y_i^{\text{actual}}]}{M} \tag{1}$$

$I[S]$ is 1 if the statement S is true and 0 otherwise, and M represents the number of classified examples.

Besides measuring the error rate, we also measured precision, recall and F-measure. In the case of multi-label classification this measures are obtained as averages over all classifier's decisions – micro-averaging and over all categories – macro-averaging.

Micro-averaged measures weight all the web pages equally, representing the averages over all the (web page, genre category) pairs. They tend to be dominated by the classifier's performance on common categories. Micro-averaged precision (π(micro)) represents the ratio of web pages correctly classified as l (TP = true positives), and all the pages correctly and incorrectly (FP = false positives) classified as l (Eq. 2). Micro-averaged recall (ρ(micro)) represents the ratio of web pages correctly classified as l, and all the pages actually pertaining to the class l (FN = false negatives) (Eq. 3). Micro-averaged F-measure (F(micro)) represents a harmonic mean of π(micro) and ρ(micro) (Eq. 4). $|L|$ represents the number of categories.

$$\pi\,(\text{micro}) = \frac{\sum_{l=1}^{|L|} TP_l}{\sum_{l=1}^{|L|}(TP_l + FP_l)} \tag{2}$$

$$\rho\,(\text{micro}) = \frac{\sum_{l=1}^{|L|} TP_l}{\sum_{l=1}^{|L|}(TP_l + FN_l)} \tag{3}$$

$$F\,(\text{micro}) = \frac{2 \times \pi\,(\text{micro}) \times \rho\,(\text{micro})}{\pi\,(\text{micro}) + \rho\,(\text{micro})} \tag{4}$$

Macro-averaged measures weight equally all the genre categories, regardless of their frequencies. They tend to be dominated by the classifier's performance on rare categories. Macro-averaged precision (π(macro)) is computed firstly by computing the precision for each category separately, and then by averaging over all categories (Eq. 5). The same procedure is used for computing the macro-averaged recall (ρ(macro)) (Eq. 6), and macro-averaged F-measure (F(macro)) (Eq. 7).

$$\pi_l = \frac{TP_l}{TP_l + FP_l}, \quad \pi\,(\text{macro}) = \frac{\sum_{l=1}^{|L|} \pi_l}{|L|} \tag{5}$$

$$\rho_l = \frac{TP_l}{TP_l + FN_l}, \quad \rho\,(\text{macro}) = \frac{\sum_{l=1}^{|L|} \rho_l}{|L|} \tag{6}$$

$$F_l = \frac{2 \times \pi_l \times \rho_l}{\pi_l + \rho_l}, \quad F\,(\text{macro}) = \frac{\sum_{l=1}^{|L|} F_l}{|L|} \tag{7}$$

6 Results and Discussion

The performances of the four classifiers induced with LibSVM and AdaBoost algorithms on the multi-class and binary data sets are presented in Table 8 and Table

9. Both multi-class classifiers correctly classified higher number of examples than the binary classifiers. To understand if this happens due to better recognition of multi-genre web pages, we broke down the correct predictions into categories according the number of actual labels and the number of correctly predicted labels (Table 10). To allow the comparisons between the multi-class and the binary classifiers, we transformed the numbers of correctly predicted labels into ratios. For example, in the case of two-genre pages there were 128 examples. The LibSVM multi-class classifier correctly predicted one of the two genres for 45 examples or the 35% of the two-genre cases, and two of the two genres for 14 examples or the 11% of the two genre cases. As can be seen from the Table 10, the removal of the examples labeled with improperly represented categories in multi-class setting (cf. Evaluation section), resulted in different number of web pages per the number of labels category.

From Table 10 it can be seen that the higher EX of the multi-class classifiers in comparison to the binary classifiers is due to better recognition of single-genre web pages (on average 10 percentage points improvement). In the case of two-genre web pages the quality of recognition was on average the same – around 10%. Because of the small number of the three-genre and four-genre web pages, we cannot make proper conclusions for more than two genres per page.

Considering other qualities of the classifier, the binary classifiers showed considerably higher precision in the comparison to the multi-class classifiers. We consider high precision as a good property of genre classifier since on the open web it is of higher importance to get precise top ten hits than to retrieve all possible hits.

7 Conclusion and Future Work

In this paper we compared two approaches to multi-label web genre classification – multi-class and binary – to understand which one can better capture the relations between genres that appear together in multi-genre web pages. To this end four multi-label classifiers were induced, two multi-class and two binary. Overall performances of both multi-class and binary classifiers are relatively low. For example, the exact match ratio is around 38% for multi-class and around 29% for binary classifiers. A potential reason is the high number of features (2,491) in comparison to the number of examples (1,539), an aspect which we intend to address in further experiments through feature selection.

Binary classifiers considerably outperformed the multi-class classifiers in precision (by around 62% when micro-averaged and around 61% when macro-averaged). However, this criterion alone is not enough to make the choice between the approaches. Further evidence showed that the differences between the two approaches were largely in different ability to correctly classify single-genre web pages. Therefore, we can conclude that under presented circumstances both approaches fail to address the issue of correct recognition of multi-genre web pages.

As part of the future work, the approaches could be tested on another data set, preferably larger and with more multi-label examples. Several other ML algorithms

could be applied. This would rule out the influences of the specific corpus and the specific ML algorithms.

References

Berners-Lee, T., Fielding, R., and Masinter, L. (1998). RFC2396: Uniform Resource Identifiers (URI): Generic Syntax. *RFC Editor United States*.

Dewdney, N., VanEss-Dykema, C., and MacMillan, R. (2001). The form is the substance: Classification of genres in text. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, pages 1–8. Association for Computational Linguistics.

Fan, R., Chen, P., and Lin, C. (2005). Working Set Selection Using Second Order Information for Training Support Vector Machines. *The Journal of Machine Learning Research*, 6:1889–1918.

Freund, Y. and Schapire, R. (1996). Experiments with a New Boosting Algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann.

Hsu, C., Chang, C., and Lin, C. (2008). A Practical Guide to Support Vector Classification. `http://www.csie.ntu.edu.tw/~cjlin`.

Karlgren, J. and Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 1071–1075. Association for Computational Linguistics.

Lim, C., Lee, K., and Kim, G. (2005). Multiple Sets of Features for Automatic Genre Classification of Web Documents. *Information Processing and Management*, 41(5):1263–1276.

Meyer zu Eissen, S. and Stein, B. (2004). Genre Classification of Web Pages: User Study and Feasibility Analysis. In *KI 2004: Advances in Artificial Intelligence*, pages 256–269. Springer.

Porter, M. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*.

Rosso, M. (2005). *Using Genre to Improve Web Search*. PhD thesis, University of North Carolina.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., and Liu, X. (2001). Genre Based Navigation on the Web. In *Proceedings of the 34th Hawaii International Conference on System Sciences*.

Santini, M. (2004). A Shallow Approach to Syntactic Feature Extraction for Genre Classification. In *Proceedings of 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*.

Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton.

Santini, M. (2008). Zero, Single, or Multi? Genre of Web Pages Through the Users' Perspective. *Information Processing and Management*, 44(2):702–737.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 26(4):471–495.

Stubbe, A., Ringlstetter, C., and Schulz, K. (2007). Genre as Noise: Noise in Genre. *International Journal on Document Analysis and Recognition*, 10(3):199–209.

Tsoumakas, G. and Katakis, I. (2007). Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Vidulin, V., Luštrek, M., and Gams, M. (2007). Using Genres to Improve Search Engines. In *Proceedings of International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 45–51.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.

Table 1: Web Genre Categories used in this paper.

Genres	Examples	Genres in (Lim et al., 2005)
Blog	blogs, diaries, time-stamped updates	N/A
Childrens'	encyclopedia for children, lyrics for children	N/A
Commercial/ promotional	homepages of institutions, organizations, political parties, institutionalized individuals; product descriptions; service descriptions; press releases	*Public homepages, Commercial homepages, Product specifications*
Community	forums, news group pages, portals with user-generated content	*Discussions, Bulletin collections*
Content delivery	download pages, image and movie galleries, games	*Image collections*
Entertainment	jokes, puzzles, horoscopes, games	N/A
Error message	custom HTTP error pages, non-HTTP errors	N/A
FAQ	faq	*FAQs*
Gateway	introductory pages, redirection pages, login pages	*Input pages**
Index	link collections, table of contents	*Link collections*
Informative	encyclopedic materials, recipes, user manuals, how-tos, lecture notes for a wide audience, informative books, biographies, discographies, filmographies	*Informative materials*
Journalistic	news, reportages, editorials, interviews, reviews	*Journalistic materials*
Official	legal materials, official reports, rules	*Official materials**
Personal	personal homepages, pages with opinions, descriptions of interests and activities	*Personal homepages, Discussions**
Poetry	poems, lyrics	*Other (informal texts)**
Pornographic	pictures and videos, stories	N/A
Prose fiction	fanfiction story, short story, novel	*Other (informal texts)**
Scientific	papers, theses, lecture notes for a specialized audience, scientific books	*Research reports*
Shopping	online stores, classified ads, price comparators, price lists	*Official materials**
User input	forms, surveys	*Input pages*

Table 2: A composition of 20-Genre Collection corpus.

Genre	No. of pages
Blog	77
Childrens'	105
Commercial/ promotional	121
Community	82
Content delivery	138
Entertainment	76
Error message	79
FAQ	70
Gateway	77
Index	227
Informative	225
Journalistic	186
Official	55
Personal	113
Poetry	72
Pornographic	68
Prose fiction	67
Scientific	76
Shopping	66
User input	84

Table 3: A set of surface features.

Features
Function words: number of occurrences of 50 most common function words in the corpus / total number of function words
Genre-specific words: number of occurrences of 321 selected content words / total number of content words
Punctuation marks: number of occurrences of 34 selected punctuation symbols / total number of punctuation symbols
Classes of words: number of named entities of the classes date, location and person / total number of words
Text statistics: average number of characters per word; average number of words per sentence; number of characters in hyperlink text / total number of characters

Table 4: A set of structural features.

Features
POS tags: number of occurrences of 36 available POS tags / total number of words
POS trigrams: number of occurrences of 1,868 selected POS trigrams / total number of POS trigrams
Sentence types: number of declarative sentences, interrogative sentences, exclamatory sentences and other sentences (in most cases list items) / total number of sentences

Table 5: A set of presentation features.

Features
Token type: number of alphabetical tokens (sequences of letters), numerical tokens (sequence of digits), separating tokens (sequences of separator characters, such as spaces and returns) and symbolic tokens (sequences of characters excluding alphanumeric and separator characters) / total number of tokens

HTML tags: number of single tags / total number of tags; number of tags belonging to a class of tags / total number of tags for 5 classes:
1. **Text Formatting:** <abbr>, <acronym>, <address>, <b>, <basefont>, <bdo>, <big>, <blockquote>, , <cite>, <code>, <del>, <dfn>, , <font>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <i>, <ins>, <kbd>, <pre>, <q>, <s>, <samp>, <small>, <strike>, , <style>, <sub>, <sup>, <tt>, <u>, <var>
2. **Document Structure:**
, <caption>, <col>, <colgroup>, <dd>, <dir>, , <dl>, <dt>, <frame>, <hr>, <iframe>,   , <menu>, <noframes>,   , <p>, , <table>, <tbody>, <td>, <tfoot>, <th>, <thead>, <tr>,
3. **Inclusion of external objects:** <applet>, <img>, <object>, <param>, <script>, <noscript>
4. **Interaction:** <button>, <fieldset>, <form>, <input>, <isindex>, <label>, <legend>, <optgroup>, <option>, <select>, <textarea>
5. **Navigation:** Counting href attribute of tags <a>, <area>, <link> and <base>

Table 6: A set of context features – URL features

Features	Description
Https	Indicates whether the scheme is https.
URL depth	Number of directories included in the path.
Document type	Described by four Boolean features, each indicating whether the document type is static HTML (document extensions html and htm), script (document extensions asp, aspx, php, jsp, cfm, cgi, shtml, jhtml and pl), doc (document extensions pdf, doc, ppt and txt) or other (the other document extensions).
Tilde	Appearance of "/ " in the URL.
Top-level domain	Described by ten Boolean features, each indicating whether the top-level domain is com, org, edu, net, gov, biz, info, name, mil or int.
National domain	Indicates whether the top level domain is a national one.
WWW	Indicates if the authority starts with www.
Year	Indicates the appearance of year in the URL.
Query	Indicates the appearance of query (?foo) in the URL.
Fragment	Indicates the appearance of fragment (#foo) in the URL.
Appearance of 54 most commonly used words in URL	Indicates the appearance of common content words in URL: *about, abstract, adult, archiv, articl, blog, book, content, default, detail, download, ebai, english, error, fanfic, faq, forum, free, fun, funni, galleri, game, help, home, index, joke, kid, legal, librari, link, list, lyric, main, member, music, new, paper, person, poem, poetri, product, project, prose, pub, public, quiz, rule, search, sport, stori, topic, tripod, user, wallpap*

Table 7: A set of context features – links.

Features
Links: number of hyperlinks to the same domain, to a different domain and containing "mailto" / total number of hyperlinks

Table 8: The performances of the two classifiers induced with the LibSVM on the multi-class and binary data sets.

	LibSVM						
Data set	EX	$\pi\,(micro)$	$\rho\,(micro)$	$F\,(micro)$	$\pi\,(macro)$	$\rho\,(macro)$	$F\,(macro)$
Multi-class	38%	0.37	0.37	0.37	0.20	0.16	0.16
Binary	29%	0.55	0.29	0.38	0.50	0.34	0.34

Table 9: The performances of the two classifiers induced with the AdaBoost on the multi-class and binary data sets.

	AdaBoost						
Data set	EX	$\pi\,(micro)$	$\rho\,(micro)$	$F\,(micro)$	$\pi\,(macro)$	$\rho\,(macro)$	$F\,(macro)$
Multi-class	38%	0.35	0.35	0.35	0.15	0.18	0.16
Binary	29%	0.68	0.32	0.43	0.71	0.34	0.44

Table 10: Correctly classified examples having one, two, three and four labels (Act. = actual labels, Pred. = predicted labels)

No-Labels		Multi-class				Binary			
Act.	Pred.	LibSVM		AdaBoost		LibSVM		AdaBoost	
1	1	169/352	48%	165/352	47%	133/353	38%	132/353	37%
2	1	45/128	35%	44/128	34%	54/146	37%	54/146	37%
2	2	14/128	11%	11/128	9%	14/146	10%	11/146	8%
3	1	2/4	50%	1/4	25%	5/13	38%	5/13	38%
3	2	0/4	0%	1/4	25%	2/13	15%	2/13	15%
3	3	0/4	0%	0/4	0%	0/13	0%	0/13	0%
4	1	N/A	N/A	N/A	N/A	0/1	0%	0/1	0%
4	2	N/A	N/A	N/A	N/A	0/1	0%	0/1	0%
4	3	N/A	N/A	N/A	N/A	0/1	0%	0/1	0%
4	4	N/A	N/A	N/A	N/A	0/1	0%	0/1	0%

Silvia Petri and Mirko Tavosanis

Building a Corpus of Italian Web Forums: Standard Encoding Issues and Linguistic Features

Abstract

This paper describes the creation of a reference corpus of nearly 1200 Web forum posts in Italian. The corpus was created evaluating and customizing a previous proposal for Xml standard encoding; a revised version of the relevant DTD is now proposed as reference for the structural features of Web forum posts and a set of correspondences, with little loss of information, is given for the TEI P5 encoding system. Preliminary results about syntactic features of the language of the posts are also included to sample the linguistic variability of this textual genre.

1 Overview

Web forums are, arguably, the most popular interactive textual genre on the web. Current Eurostat surveys show that up to 50% of the citizens of some states of the European Union have posted at least a Web message in the year preceding the interview: this posting is undoubtedly often a *forum* posting. However, few studies (such as Light and Rogers 1999) have dealt with the linguistic or textual features of the forums or even with more basic facts such as their diffusion.

Moreover, there is widespread variety even in the name of the genre. Describing Web forums, secondary literature calls them *message boards*, *discussion boards*, *discussion groups*, *conversations*, *chatgroups*, *newsgroups* and so on (as for classification issues for Web texts, see also Rehm et al. 2008). This variation also seems to push towards the grouping in the literature of many textual genres we find scarcely related from every point of view: Usenet newsgroups of the 1990s lack many of the features now common in Web forum interfaces, and so on. Typical of this attitude are syntheses such as Crystal (2006, pp. 134-177) where a single chapter devoted to "The language of chatgroups" describes both "asynchronous" and "synchronous groups", with a few lines of specification.

In this paper we will instead deal with a very specific textual genre: asynchronous conversations collected in threads and managed by a particular kind of Web site. Many Web forums allow more than a single way of interaction and messages can be posted on them by various means (Web interface, e-mail and so on). We will see in §§ 4-5 that this seems to have linguistic consequences more relevant than those connected to, say, the topic of the forum, hinting to the need to make subtler distinction between subgenres.

We did not take into account, then, traditional newsgroups or mailing lists if they don't have a web interface allowing users not only to read past conversations but also to write new messages. On the other hand, we did consider as possible sample material every kind of forum where:

- messages are archived on the Web in antichronological order, and,
- a Web interface allows users to compose posts on-line and to publish them on the system.

Our description of Web forums (or simply *forums*, in the following sections of the paper) will then be relative to this kind of publication. We feel that the genre features of Web forums are both specific enough to be described on their own a#nd varied enough to require in-depth discussion. In order to verify this feeling, we then decided to create a reference corpus of Italian forums.

2 Selection of Forums

Forums can undoubtedly be dedicated to different topics and can have very different features and very different language choices. Moreover, their features seem only partially standardized. In order to sample this variety, we choose to start the corpus construction with four Italian forums of different character and managed with different software tools, including the two packages which together seem to cover more than half of the market, phpBB and vBulletin:

Accademia della Crusca Formal discussions about Italian language (forum closed in 2005 and managed with phpBB software) (*http://forum.accademiadellacrusca.it/phpBB2/index.php*)

ADI – Associazione dei dottorandi e dottori di ricerca italiani Discussions about work issues of Ph.D,. students and Ph. D. holders (forum managed with phpBB software; messages can be sent to the forum both through a Web interface and a mailing address and are distributed both through the Web and a mailing list) (*http://www.dottorato.it/forum/index.php*)

HTML.it Technical forums about HTML and related languages (forum managed with house-developed software) (*http://forum.html.it/forum*)

NGI Social forum, gathering players of an online role-playing game (forum managed with vBulletin software) (*http://gaming.ngi.it/forum/forumdisplay.php?f=501*)

This selection is absolutely arbitrary: it is useful to stress that to our knowledge there are absolutely no reliable data about the distribution of forums to use as a starting point to build a "representative" corpus (see also Leech 2007 for a critical appraisal of the current corpus "representativeness"). Nor is the sheer number of forums known, even if there are many hundreds (or thousands) of them in the Italian Web alone. Random searches allow to find Italian forum sites with as many as 509,962 registered users, 21,838,712 messages and 1,374,969 threads each. As for the world scene, the BoardTracker site boasts the

inclusion in its database of nearly 38,000 forums and of 59 millions of threads (Petri 2008, 6-8, 13), but actual figures could be much higher. The Gaia Online forum alone declares on its home page (19.9.2008) to host "**1,412,466,251** articles posted with **14,646,253** registered users". Anyway, at the moment there are also no reliable estimates about the most popular forum topics (games? general discussions?) or the most used languages in forum posts.

Forums 1 and 2 were then chosen, according to previous knowledge, as potential samples of a more formal language and style of communication; forums 3 and 4 as potential samples of less formal language. On the whole, we gathered from the four forums a grand total of 1186 posts (the size limit being imposed mainly by work constraints: see § 3), with different criteria.

From the first two forums were randomly selected four threads which were composed by 25 posts or more. The posts were encoded integrally (see § 2 for mark-up description) and in the original order. The thread encoding proceeds then from the newest to the oldest post.

Instead, from the HTML.it and NGI forums has been extracted a bigger sample: 50 threads of 10 posts each. This choice was based upon the hypothesis that these forums would have represented better the less formal language levels which seem more common on the Web even if, given the status of knowledge about this textual genre, there is no hard evidence to support this idea, which relies only on common sense (and can be disproved by future research).

Also in this case, the procedure of selection began from the first page of the forum site and has gone on turning back in the chronology of the forum to find new material: since this work required some weeks, every day the selection began again from the first page and from the "most updated" threads it hosted. For these forums too the threads were chosen according to the number of replies they included, discarding threads with less than 10 replies.

In the revision phase, a few of the posts were then discarded due to mistakes in selection. At the end of the work, the encoded posts were then 1186 instead of 1200.

As expected, interface features and style of writing seem to vary according to the forum. The selected posts of the forum of the Accademia della Crusca often offer high-handed discussions of language in a very formal tone. The NGI forum posts, instead, sport a variety of features of neostandard Italian (Berruto 1987). More precise descriptions, however, required the establishment of a frame of reference.

3 The Mark-up

Italian forums, as well as international ones, apparently are not gathered by standard sites hosting thousands of them, as is the case of blogs (Tavosanis 2007). Forums, instead, seem often created as additional features in many sites, and, even if the two products recalled in § 2 seem dominant, this creation is mediated by a wide variety of software (Wikipedia 2008). A widespread Html re-

ference framework does not exist, and the source code of the forum pages follows no standard. Those features of forums pose of course many limits to the performance of search engines and crawling systems (Limanto et al. 2005, 978). In order to compare different forums we have then to reduce them to a common encoding for further processing.

This work has tried to verify the conditions for a standard Xml encoding. Its starting point has then been the standard proposed by Claudia Claridge (2007, p. 94-97). This standard was then modified according to the actual features of the forums encoded, as they emerged during this research.

3.1 Starting Point

To encode and interchange materials, an Xml standard of mark-up was proposed by Claridge (2007); we will refer from this point to this model simply as "Claridge". It includes the following elements and attributes:

```
<forum> (root element)
<thread>
<message> with attributes "topic", "no" and "ad"
<person> with attributes "gender" and "desc"
<mbinfo> (message board-related information) with attributes "joined",
"posts", "avday", "greats" and "warnings"
<place> with attribute "desc"
<time>
<sig> (marks an automatic attachment to any message a given writer
sends; also called signature)
<body> (the content of the message)
<p> (the paragraphs in the body)
<quote> (marks the quotations)
<visual> (marks the presence of "graphic elements with emotive/interac-
tive meaning" (Claridge 2007, p.97)) with attribute "meaning"
<gap> (marks "purely additional, decora-
tive material" (Claridge 2007, p.97))
```

This structure was applied to the Html of the threads selected according to the procedure described in § 2. Of course, this required a heavy work of cleaning, since the typical Html code of an average post looks in this way (the example is drawn from the third selected thread of the HTML.it forum):

```
<TD style="MIN-HEIGHT: 280px; WIDTH: 200px" vAlign=top bgColor=#f0f0f0
wrap><A name=post11412788></A><SPAN class=big><B>Myaku</B></
SPAN><BR><SPAN class=little>Utente di HTML.it</SPAN><BR><BR><IMG
alt="" src="Thread3 - Iframe_file/hetfield2.jpg" border=0><BR><BR><SPAN
class=little><P class=postbit_dati>Registrato il: Nov 2006</P><P
class=postbit_dati>Provenienza: </P><P class=postbit_dati>Messaggi:
2850</P><BR><P class=postbit_dati>ICQ : </P><P class=postbit_
dati>MSN : chiedere in privato</P><P class=postbit_dati>Skype : <A
```

```
href="skype:chiedere%20in%20privato?chat">chiedere in privato</
A></P></SPAN><BR></TD><TD style="MIN-HEIGHT: 280px; WIDTH: 712px"
vAlign=top bgColor=#f0f0f0><SPAN class=little><B>Re: Iframe</B></
SPAN> <DIV class=corpopost onfocus=this.blur();><DIV class=head_
citazione>Citazione:</DIV> <DIV class=citazione><STRONG>Originariamen
te inviato da mayorca </STRONG><BR>Ho utilizzato Dreamweaver CS3, non
vorrei che avesse inserito qualche tag che mi blocca le pagine.<BR></
DIV><BR><BR>se non ci fai vedere il codice, mi sa che anche a tir-
are a indovinare la vedo dura <IMG alt="" src="Thread3 - Iframe_
file/smile.gif" border=0><BR><BR>ps. consiglio: evita i frames/
iframes se puoi<BR><BR><IMG alt="" src="Thread3 - Iframe_file/ciao.
gif" border=0></DIV><P><SPAN class=norm><BR>_____<BR><FO
NT style="LINE-HEIGHT: 150%" size=1><STRONG><A href="http://www.sysu-
niverse.net/" target=_blank>grafica &amp; web</A> | <A href="http://
www.thinksy.altervista.org/" target=_blank>blog</A> -- no pvt tecni-
ci -- </STRONG><BR>AVVISO: l'aiuto è gratuito, la "pappa pronta" si for-
nisce solo su preventivo.<BR><FONT style="COLOR: red">...USATE IL TAG
# quando vi chiedo di postare il codice!!!!</FONT></FONT></SPAN></P>
```

The differences in encoding made unavoidable to clean and to re-encode the corpus largely by hand. We are aware of current automatic cleaning practices such as those documented in the CleanEval competition, but these kind of approaches encountered many difficulties and we judged that manual encoding was the only way to have satisfactory results in the time frame available for the work. At the end of the process, it became evident that it was in fact possible to encode the posts using the Claridge structure, but also that some features could be represented in a more satisfactory way by making a few changes to the model.

3.2 Changes to the Original Structure and Final Mark-up

In our revision, the XML tree remains unchanged, with the root element <forum> that contains some elements <thread>, composed by elements <message>.

The main element <forum> has the new attributes "name" (the name of the corresponding forum), "section" (the section analyzed), "url" (including the web address of the forum), and "software" (indicating the software used for the creation and management of the forum).

The element <thread> has a "cod" attribte (to identify the thread univocally with two letters of the alphabet), and a "topic" with the argument of that thread.

The <message> element has now the following attributes:

- "title" (corresponding to "topic" in the original model)
- a progressive number ("no")
- an indication allowing to correlate directly the message with another of which it includes quotations ("ad").

In every message we encoded information about the writer: as in the Clar-

idge mark-up, we have the <person> element, but we kept only the attribute "desc", a description of the rank reached from that sender in the forum, or (like in the NGI forum) a phrase or a title inserted from the sender to describe him- or herself, while the foreseen attribute "gender" has never found explicit correspondence, either in the forum or in the user profile.

Other data about the writer are fitted in the element <mbinfo> that includes the attributes "joined", "posts" (as in Claridge), "place" (that was an element by itself in the original mark-up but that has been inserted included here because it refers to the sender and not to the message), and other particular attributes such as "ICQ", "MSN" and "Skype".

In the corpus it has been found no information to instantiate the attributes "avday" ("the average posting rate per day" (Claridge, p. 96), present only in each user profile and of small interest for this kind of work), "greats" ("an evaluative category of "great" message" (Claridge, p. 96)) and "warnings" ("the number of warnings a sender has received for violating netiquette or forum rules" (Claridge, p. 96)), which were then dropped.

Every message contains the element <time> (as in Claridge) that indicates the temporal coordinates of the message writing, with the attribute "edited", in the case of a later editing by the writer; in two cases there was an indication of the reason of the editing, marked in the attribute "reason".

For the central part of the message, the element <body>, the original HTML structure has been kept, with , , <br/>, <i>, <b> etc. The style element indicating the font color has been marked with the element <color>, while the HTML <a href> has been changed in the Xml <link url> in order to be more understandable. A new element, named <object>, has been introduced to mark external objects in the message, such as video clips.

It has also been added an element <s> that is useful to the scope of the research allowing to encode the sentence boundaries of the text.

The element <quote> is kept as in the original encoding. The frequent element <sig> has been separated from the <body> of the message.

Regarding the encoding of graphic elements included in the message, this work follows Claridge's indications: the presence of images is marked with element <gap> while an emoticon is described in the element <visual> with attribute "desc" to encode its name and meaning.

At the end of the work, the post seen in § 3.1 is then encoded in this way:

```
<message title="Re: Iframe" no="02" ad="01">
<person desc="Utente di HTML.it">Myaku<gap/></person>
<mbinfo joined="Nov 2006" posts="2850" MSN="chiedere in privato"
Skype="chiedere in privato"/>
<time>18-02-2008 12:33</time>
<body>
<div class="corpopost"><quote>Citazione:<b>Originariamente invia-
to da mayorca </b><br />Ho utilizzato Dreamweaver CS3, non vor-
```

```
rei che avesse inserito qualche tag che mi blocca le pagine.<br /></
quote><br /><br /><s>se non ci fai vedere il codice, mi sa che anche
a tirare a indovinare la vedo dura <visual desc="Smile"/></s><br /><br
/><s>ps. consiglio: evita i frames/iframes se puoi</s><br /><br /><vi-
sual desc="ciao"/></div><br/>_____<br/></body><sig><b><link
url="http://www.sysuniverse.net">grafica &amp; web</link> | <link
url="http://www.thinksy.altervista.org">blog</link> -- no pvt tecnici --
</b><br />
: l'aiuto è gratuito, la "pappa pronta" si fornisce solo su
preventivo.<br />
<color n="red">...USATE IL TAG # quando vi chiedo di postare il codi-
ce!!!!</color>
</sig>
</message>
```

The final version of the encoding was then archived for future work and published in our web pages. The current DTD is published as Appendix at the end of this paper.

3.3 TEI P5 Encoding

In order to make easier to reuse the corpus, we established also a set of correspondences of our Xml elements and attributes with TEI P5 elements and attributes. A few attributes considered of little use to researchers were deleted, while the new elements <byline> and <closer> were enclosed. This is the final list of correspondences:

Claridge revised	TEI P5
<forum> attributes: the values of "name", "section", "url" and "software" are enclosed as text in the <title> element (child of <sourceDesc>)	<teiCorpus>, enclosing <TEI> and <teiHeader>
<thread> attributes: "cod" values become values of "xml:id" of <div1>; "topic" values are enclosed as text in <head> (child of <div1>)	<div1 type="thread">
<message> attributes: "no" values become values of "n" of <div2>; "ad" values are suppressed; "title" values are enclosed as text in <head> (child of <div2>)	<div2 type="message">
<person> attributes: "desc" values are suppressed	<persName>

<mbinfo> attributes: “place” values are enclosed as text in <placeName>; “joined” values are enclosed as date in <date type=”jointime”>; “posts” values are enclosed as text in <num type=”msgnumber”>; attributes “MSN”, “ICQ” and “Skype” are suppressed	
<time> attributes: “edited” and “reason” values are suppressed	<time>
<body>	<p>
<s>	<s>
	suppressed
	suppressed
<td>	suppressed
<quote>	<seg type=”quotation”> enclosing <quote>
<sig>	<closer> enclosing <signed>
<link> attributes: “url” values become values of “target” of <ref>	<ref>
<u>	<emph rend=”underline”>
<i>	<emph rend=”italic”>
<b>	<emph rend=”bold”>
<color> attributes: “n” values are suppressed	<emph rend=”color”>
 	<lb>
<object> attributes: “desc” values are suppressed	<binaryObject>
<gap>	<graphic>
<visual> attributes: “desc” values are enclosed as text in <figDesc> (child of <figure>)	<seg type=“iconic“> enclosing <figure>

As for the relevant Schema, it was generated through the online tool Roma (http://www.tei-c.org/Roma/) adding to the standard set of TEI modules (Core, Tei, Header, Textstructure) the following optional modules:

- Analysis
- Corpus
- Namesdates
- Figures
- Linking

The corpus was then transformed using an XSL-T stylesheet. The output was a valid TEI P5 file. In this version, the post seen in § 3.1 is encoded in this way:

```
<div2 type="message" n="02">          <head>Re: Iframe</head>       <by-
line>
        <persName>Myaku</persName>
        <time>18-02-2008 12:33</time>
        <placeName></placeName>
        <date type="jointime">Nov 2006</date>
        <num type="msgnumber">2850</num>
        </byline>
        <p><seg type="quotation">
        <quote>Citazione:<emph rend="bold">Originariamente inviato da
mayorca </emph><lb/>
Ho utilizzato Dreamweaver CS3, non vorrei che avesse inserito qual-
che tag che mi blocca le pagine.<lb/></quote></seg><lb/><lb/><s>se non
ci fai vedere il codice, mi sa che anche a tirare a indovinare la vedo
dura <seg type="iconic"> <figure> <figDesc>Smile</figDesc> </figure></
seg></s><lb/><lb/><s>ps. consiglio: evita i frames/iframes se puoi</
s><lb/><lb/><s><seg type="iconic"><figure>
<figDesc>ciao</figDesc></figure></seg></s><lb/>_____<lb/></
p>
        <closer><signed><emph rend="bold"><ref target="http://www.
sysuniverse.net">grafica &amp;amp; web</ref> | <ref target="http://
www.thinksy.altervista.org">blog</ref> -- no pvt tecnici -- </
emph><lb/>AVVISO: l'aiuto è gratuito, la "pappa pronta" si fornisce solo
su preventivo.<lb/><emph rend="color">...USATE IL TAG # quando vi chiedo
di postare il codice!!!!</emph></signed> </closer>
<div2>
```

4 Linguistic Data Extraction

In addition to the validation of the structural elements, we used the corpus to evaluate some linguistic features of the texts. The starting numbers are:

- Number of posts: 1186
- Number of word forms: 150115
- Average number of word forms per message: 125

The number of words per message seems of particular interest:

Forum	*Words / message ratio*
Accademia della Crusca	111
ADI	470
HTML.it	108
NGI	75
Total	125

A second step was the encoding and evaluation of sentences. Many of the messages included non-standard constructions, especially from the punctuation point of view. In particular, ellipses (three or more full stops) and line breaks are used both for in-sentence pauses and as end-of-sentence marks. A sample of this can be seen in the following section, where the HTML tag <br/> is used to mark the limits of two sentences:

```
<s>evita JS finché puoi</s><br />
<s>evita frame e iframe finché puoi</s><br />
<s>usa le inclusioni asp o php per le parti co-
muni (testata, menu, footer...)</s>
```

This situation makes strict human supervision mandatory in order to have reliable sentence counts. At the end of the process, then, the grand total for the whole corpus was:

Sentences: 3435
Average number of sentences per message: 2.9

Those numbers can now be compared with those given by the few relevant studies for other electronic text types. Moreover, we can start clarifying the differences between different textual subgenres.

The average number of sentences in a message, in fact, seems more or less stable between the different forums:

Forum	*Sentence / message ratio*
Accademia della Crusca	3.1
ADI	4.9
HTML.it	2.9
NGI	2.4

The same is true for the word / sentence ratio (which is also surprisingly high for Italian prose standards):

Forum	Word / sentence ratio
Accademia della Crusca	35
ADI	100
HTML.it	37
NGI	31
Total	43

From the linguistic and textual point of view it is a little surprising to note that highformality forums such as those of the Accademia della Crusca are only

slightly more articulated from the syntactic point of view than supposedly low-key forums such as those dealing with on-line gaming.

5 Use of Emoticons

On a different level, we investigated also the role of emoticons. Even in this case we found a surprisingly stable distribution:

Forum	% of emoticons to messages
Accademia della Crusca	48
ADI	13
HTML.it	48
NGI	46
Total	44

About the emoticons, we found out some characteristics that could indicate a first categorization of their functions and use:

1) Can be substituted by adverbs (13 cases)
- example from HTML.it: "<s> Allora ho provato ad usare il metodo get al posto di post… funziona <visual desc="Mmmm… strano… molto strano"/> </s>"

2) Can be substituted by verbal expressions (3 cases)
- example from NGI: "<s> avuti 3 rogue, 1 prete e ora ne sto rollando n'altro <visual desc="Love"/> </s>"

3) Emphasize graphically the meaning of the words surrounding them (143 cases)
- example from Accademia della Crusca: "<s> Allora continuiamo a indagare <visual desc="Rolling Eyes"/> </s>"

4) Express the attitude of the writer (270 cases)
- example from ADI: "<s> A tuo piacimento guarda…. tanto l'effetto non cambia, ahimè sono andata
 <visual desc="Smile"/> </s>". This is the original function of emoticons but, evidently, not the only one.

5) Can be substituted by imprecations (23 cases)
- example from HTML.it: "<s> Ho appena sentito il cliente, gli ho fatto fare dei test, non funziona <visual desc="Mannaggia li pescetti"/> </s>"

6) Form a phrase by themselves (82 cases)
- example from HTML.it: "<s> <visual desc="ciao"/> </s>"

7) Are used for their graphical appearance in the composition of a word (1 case)
- example from NGI: "<s> Eventualmente sono disponibile anche <visual desc=":v"/>ome riserva nel 3v3 o 5v5. </s>"

6 Data Evaluation

We suspect that the most conspicuous variation in our data, the highest complexity of the messages of the ADI forum and their relative lack of emoticons, is related to the nature of the writing medium. As anticipated in § 2, the ADI

forum is the only source in our corpus including messages sent through a mailing list. The system does not allow to check if a message is composed with mail software or with forum interface; we suspect however that many messages are composed as e-mail messages and that the strong difference between the two mediums explains the difference in message and phrase length. Even the relative lack of emoticons could be easily explained in this way, since mail software usually does not display ready-to-use emoticons, while forum interfaces display them.

In order to sample the internal variety of forums, we also selected a single thread of the Accademia della Crusca forum. The thread ("Auguri") was completely dedicated to the exchange of well-wishing messages for Christmas and New Year's Eve and was then of a very different character front of linguistic discussion threads. The counts for this single thread are:

sentence / message ratio: 2.5
word / sentence ratio: 25
message length in words: 63
% of emoticons to messages: 68

The difference between those values and those of the Accademia della Crusca forum as a whole is noteworthy, and it is consistent with the general idea of how such a thread can be different from a scholar discussion. However, it is also useful to note that this difference is not particularly striking in quantitative terms.

7. Conclusions and future work

Even if the Xml encoding of forums for research purposes is rare, we feel that our work displays the feasibility and utility of such a practice. The corpus described is now available through our institutional web pages: we hope that the collected materials can be used as term of comparison for further analyses of forums, especially from the linguistic point of view.

Of particular interest seems the stability in phrase and message length across different topics and different social contexts. This hints to a dominant role of tool and genre, more than topic and situation, in the creation of Web texts. Further analyses could be able to better describe this situation.

Bibliography

Berruto, G. (1987). Sociolinguistica dell'italiano contemporaneo. Roma: Carocci.

Claridge, C. (2007). „Constructing a corpus from the web: message boards." In: Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) (2007). Corpus Linguistics and the Web. Amsterdam; New York: Rodopi, 87-108.

Crystal, D. (2006). Language and the Internet. Second edition. Cambridge University Press: Cambridge (UK).

Leech, G. (2007). „New resources, or just better old ones? The Holy Grail of representativeness." In: Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) (2007). Corpus Linguistics and the Web. Amsterdam; New York: Rodopi, 133-150.

Light, A. and Rogers, Y. (1999). „Conversation as Publishing: the Role of News Forums on the Web." In: Proceedings of the 32nd Hawaii International Conference on System Sciences – Journal of Computer-Mediated Communication, 4, 4.

Limanto, H. Y. et al. (2005) „An Information Extraction Engine for Web Discussion Forums." In: International World Wide Web Conference archive. Special interest tracks and posters of the 14th international conference on World Wide Web table of contents. Chiba, Japan. New York: Association for Computing Machinery, 978-979.

Petri, S. (2008). „I forum italiani: analisi linguistica e problemi di codifica." Master Degree thesis, Università di Pisa.

Rehm, G. et al. (2008). „Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems". In: Proceedings of the LREC 2008.

Tavosanis, M. (2007). „Juvenile Netspeak and subgenre classification issues in Italian blogs." In: Rehm, G. and Santini, M. (2007). Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing. Shoumen: Incoma, 37-44.

Wikipedia. (2008). Comparison of Internet forum software. Ad voc. (***http://www.wikipedia.org***).

Appendix: the full DTD

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT forum (thread+)>
<!ATTLIST forum
                name CDATA #IMPLIED
                section CDATA #IMPLIED
                url CDATA #IMPLIED
 software CDATA #IMPLIED>
<!ELEMENT thread (message+)>
<!ATTLIST thread
                cod CDATA #IMPLIED
                topic CDATA #IMPLIED>
<!ELEMENT message (person, mbinfo,  time, body, sig?)>
<!ATTLIST message
                title CDATA #IMPLIED
                no NMTOKEN #IMPLIED
                ad NMTOKEN #IMPLIED>
<!ELEMENT person (#PCDATA | gap | link)*>
<!ATTLIST person
                desc CDATA #IMPLIED>
<!ELEMENT mbinfo EMPTY>
<!ATTLIST mbinfo
                 joined CDATA #IMPLIED
                 posts CDATA #IMPLIED
```

```
                place CDATA #IMPLIED
                bowling CDATA #IMPLIED
                ICQ CDATA #IMPLIED
                MSN CDATA #IMPLIED
                Skype CDATA #IMPLIED>
<!ELEMENT time (#PCDATA)>
<!ATTLIST time
              edited CDATA #IMPLIED
              reason CDATA #IMPLIED>
<!ELEMENT body (#PCDATA | span | td | br | s | div | quote)*>
<!ELEMENT sig (#PCDATA | br | link | i | color | u | b | visual)*>
<!ELEMENT s (#PCDATA | color | br | i | gap | visu-
al | b | link | u | span | td | s |div | quote)*>
<!ELEMENT span (#PCDATA | b | i | u | br | s | vi-
sual | gap | color | link | quote)*>
<!ATTLIST span
               class CDATA #IMPLIED>
<!ELEMENT td (#PCDATA | br | i | visual | span | b | td | u)*>
<!ATTLIST td
               class CDATA #IMPLIED>
<!ELEMENT br EMPTY>
<!ELEMENT div (#PCDATA |  b | i | u | br | s | visu-
al | gap | link | div | span | object | quote)*>
<!ATTLIST div
              class CDATA #IMPLIED>
<!ELEMENT quote (#PCDATA | b | br | visual | link
| i | gap | span | quote | u | color)*>
<!ELEMENT link (#PCDATA | i | b | color | gap | u)*>
<!ATTLIST link
               url CDATA #IMPLIED>
<!ELEMENT i (#PCDATA | br | visual | color | u | b | link | div)*>
<!ELEMENT b (#PCDATA | link | color | br | i | u | visual)*>
<!ELEMENT u (#PCDATA | br | b)*>
<!ELEMENT color (#PCDATA | b | br | link | vi-
sual | color | u | i | quote)*>
<!ATTLIST color
               n CDATA #IMPLIED>
<!ELEMENT gap EMPTY>
<!ELEMENT visual (#PCDATA)>
<!ATTLIST visual
               desc CDATA #IMPLIED>
<!ELEMENT object EMPTY>
<!ATTLIST object
              desc CDATA #IMPLIED>
```

Marina Santini, Serge Sharoff

Web Genre Benchmark Under Construction

The project presented in this article focuses on the creation of web genre benchmarks (a.k.a. web genre reference corpora or web genre test collections), i.e. newly conceived test collections against which it will be possible to judge the performance of future genre-enabled web applications. The creation of web genre benchmarks is of key importance for the next generation of web applications because, at present, it is impossible to evaluate existing and in-progress genre-enabled prototypes. We suggest focusing on the following key points: 1) propose a characterisation of genre suitable for digital environments and empirical approaches shared by a number of genre experts working in automatic genre identification; 2) define the criteria for the construction of web genre benchmarks and draw up annotation guidelines; 3) create web genre benchmarks in several languages; 4) validate the methodology and evaluate the results. We describe work in progress and our plans for future development. Since it is sometimes difficult to anticipate the difficulties that will arise when developing a large resource, we present our ideas, our current views on genre issues and our first results with the aim of stimulating a proactive discussion, so that the stakeholders, i.e. researchers who will ultimately benefit from the resource, can contribute to its design.

1 The Concept of Genre

The concept of genre is hard to agree upon. Many interpretations have been proposed since Aristotle's *Poetics* without reaching any definite conclusions about the inventory or even principles for classifying documents into genres. Some studies put the number of genres to 2,000 (Görlach, 2004) or even 4,500 (Adamzik, 1995). Additionally, the lack of an agreed definition of what genre is causes the problem of the loose boundaries between the term 'genre' with other neighbouring terms, such as 'register', 'domain', 'topic', and 'style'. The inventory of genres can be based on linguistic theories or 'folksonomies', i.e. labels used by users (Rosso and Haas, ming). For instance, users are confident with a term like *novel*, whereas linguistic researchers may prefer functional terms, like *recreation* to indicate a wider range of texts aimed at recreational reading.

Recently, definitions of genre have been adapted to the new digital environments, e.g., (Yates and Orlikowski, 1992; Erickson, 1999; Toms and Campbell, 1999; Beghtol, 2001; Heyd, 2008; Bateman, 2008). Undoubtedly, the situation on the web is more difficult than in the offline world, because the web is new, genres are fluid, web documents are very often characterised by a high level of hybridism, by the fragmentation of textuality across several documents, by the impact of technical features such as

hyperlinking and posting facilities. Nevertheless, as stressed by Karlgren (2005) the term 'genre' is established and generally understood, at least intuitively, by web users, and it is currently employed in many web-based real-world environments. For instance, online bookshops, like Amazon, organise their catalogues by genre, even if their genres are not defined in a systematic way, e.g., in addition to proper genres the Amazon list contains subject labels, like Arts, Computing or Science[1].

At present, many researchers in different fields are working with genres of electronic documents, such as FAQs, e-shops, home pages, or conference websites in order to better satisfy users' needs in a number of different application areas, such as information retrieval, e.g., (Stamatatos et al., 2000; Meyer zu Eissen and Stein, 2004), digital libraries, e.g., (Rauber and Müller-Kögler, 2001; Kim and Ross, ming), and information extraction, e.g., (Maynard et al., 2001; Gupta et al., 2006). Arguably, genre is a fundamental concept in information management and definitely deserves in-depth investigations.

Genre-Enabled Prototypes Attempts at automatic genre identification of the Brown Corpus[2] start with (Karlgren and Cutting, 1994; Kessler et al., 1997). The first prototype of a genre-enabled application for the web was created in 1998 (Karlgren et al., 1998) (see DropJaw below). More recently, a genre add-on that can be installed on to a general-purpose search engine (namely Mozilla Firefox) has been completed at Bauhaus University Weimar, Germany (Stein et al., ming) (see WEGA below). In both cases, these applications could not and cannot be fully evaluated because of the absence of web genre benchmarks enabling the objective assessment of their effectiveness. Yet, the design and the construction of genre-enabled prototypes show the potential of genre in real-world applications.

All in all, four prototypes have been described and documented, namely: DropJaw, Hyppia, X-Site and WEGA.

DropJaw (for English) – Karlgren and co-workers (Karlgren et al., 1998) built a fully functional prototype system, DropJaw, to experiment with iterative search on the web. DropJaw bases its searches for web documents on terms entered by the user, as in a traditional system. However, rather than producing ranked lists of output based on term occurrence, DropJaw displays the distribution of the resulting set over two dimensions: dynamically generated topical clusters and document genres. The two-dimensional document space is displayed on a work board or matrix for further user processing.

Hyppia (for English) – The Hyppia demo allows news articles to be filtered and searched based on genre information. The genre classes in this demo are considered to be "whether a document is subjective or objective" (Finn et al., 2002; Finn and Kushmerick, 2006). (Dimitrova and Kushmerick, 2003) contributed to the Hyppia project by showing how shallow text classification techniques can be used to sort the documents returned by web search engines according to genre dimensions, such as the degree of expertise assumed by the document, the amount of detail presented, or whether the document reports mainly facts or opinions.

[1] http://www.amazon.co.uk/Books-Categories/b?ie=UTF8&node=1025612

[2] http://en.wikipedia.org/wiki/Brown_Corpus

X-SITE (for English) – X-Site is a search system designed and implemented to test the practical value of making use of task-genre relationships in a real-life work environment (Freund, 2008). X-Site was implemented as an extension to MultiText, a pre-existing indexing and retrieval engine (further details about MultiText can be found in (Freund, 2008)). X-Site makes use of three contextual components in addition to the basic search engine functionalities, namely 1) a genre classifier, which uses machine learning methods; 2) a task profile, which is composed of a work task and an information task; and 3) a task-genre association matrix, which specifies the relationships between task taxonomies and genre taxonomies.

WEGA (for English and German) – While X-Site has been devised for professionals (namely software engineers) who can exploit the concept of genre to rapidly find information that is task-appropriate, situationally-relevant and mission-critical for their job, WEGA (an acronym that stands for WEb Genre Analysis), (Stein et al., ming) has been designed for the web and for common web users. WEGA is an add-on that superimposes genre labels a few seconds after the result list is returned by a general-purpose search engine, namely Mozilla Firefox.

These prototypes show that genre-enabled systems are feasible and that genre classes can help improve productivity in the workplace (in the case of X-Site) and offer additional hints about the nature of the web pages listed in the search results (in the case of DropJaw, Hyppia and WEGA). Additionally, a number of patent applications has been submitted in the United States by XEROX Corporation on the basis of work from (Kessler et al., 1997)[3].

The design and construction of genre-annotated resources is also very timely since genre-enabled applications are a hot topic in current research, e.g., (Mehler et al., ming). The required next step is to provide evaluation resources to test these applications.

In this article we outline a project for the creation of web genre benchmarks, against which it will be possible to judge the performance of genre-enabled web applications.

2 Existing Corpora and Problems

Web genre benchmarks are still missing because their design and construction is difficult. So far, many national and 'ad-hoc' corpora have been built to represent the language, but very few large corpora indicate the genres of the documents they include, and when they do, classifications are not consistent. For example, there are several competing genre-related classifications available in the British National Corpus (BNC), such as the publication medium (book, periodical, etc), audience level, as well as a set of 70 labels called *genres*, such as 'academic texts in social sciences' (Lee, 2001). The genre attribute was included in a few collections used in information retrieval (TREC HARD 2003 and 2004, or TREC-2006 Blog Track), but the set of genres proposed was either debatable, e.g. the 'reaction' genre in TREC HARD 2003, or limited to a single genre, e.g. the BLOG genre in TREC-2006 Blog Track.

[3] For instance, see `http://http://www.patentgenius.com/patent/6973423.html`

Not happy with the genres included in these kinds of corpora, many researchers have created their own genre collections with their own inventories of genre categories. Some researches have created a hierarchy where super-genres are broken down to different medium-level genre classes, e.g., (Stubbe and Ringlstetter, 2007). Others have used more general categories such as the functional styles of the Russian linguistic tradition derived from the Prague Linguistic Circle, e.g. *everyday* or *journalistic* (Braslavski, ming), or functional classes derived from the corpus-linguistic tradition, e.g. *instructional* or *recreation* (Sharoff, ming).

While many current genre collections have the individual web pages as unit of analysis, another line of genre research focuses on genre classes at web site level. For instance, Symonenko (2007) identifies genre-like regularities in the content structure in commercial and educational websites; Rehm (2002) analyses the genre of academic personal homepages, while Mehler et al. (2007) focuses on city websites, conference websites, and personal academic homepages.

In this blossoming of genre classes and genre corpora assembled with interest-specific criteria, a practice has been established very recently, namely the testing of classification models over several existing web genre collections. This *cross-testing* technique has been adopted by Santini (ming), Kim and Ross (ming), Kanaris and Stamatatos (2009) and others. This practice represents a step forward, but only partially addresses the issues underlying the need for a more objective assessment of genre classes. Table 1 shows publicly available genre collections that have been used for cross-testing[4]. As a matter of fact, existing genre collections have been built without the ambition of being genre benchmarks. They have been created with subjective criteria following interest-specific goals. Consequently they do not have the requirements for being a "reference" or a "standard", because reference corpora, like the British National Corpus or the American National Corpus, have been built on a large consensus and based on principled criteria such as representativeness or balance.

Existing genre collections leave a number of issues unanswered. For instance, we do not know in which way they represent the genre population on the web (see the number of genres in column 3, Table 1). Additionally, they are not large enough to ensure any representativeness of individual genres, since each genre is represented by a small number of documents (from 10 to 200). Without large and comprehensive web genre benchmark spawned by a wide and comprehensive discussion on genre, it is hard to compare different empirical approaches and evaluate progress. For instance, how does the list of 298 genre labels collected by Crowston et al. (2009) compare against the set of eight genres used in the KI-04 corpus used by Stein et al. (2009)? Is the 96% accuracy reported by Kim and Ross (2009) better than the 86% accuracy obtained by Sharoff (2009)? These are the questions for which we need to find answers with the construction of large and reliable genre resources. The ultimate goal is then to

[4]These collections are all linked from the WebGenreWiki `http://purl.org/net/webgenres`. It is worth pointing out that the SANTINIS corpus also contains pages without genre annotation (considered as "noise" for the purposes of machine learning), and KRYS-I collection contains PDF documents, and not HTML pages like all the other collections listed in Table. The WebGenreWiki also contains other resources and additional discussions on genre-related issues.

Table 1: Existing web genre collections publicly available

Source	# pages	Genres
KI-04 (Meyer zu Eissen and Stein, 2004)	1205	8
SANTINIS (Santini, ming)	2480	11
I-EN (Sharoff, ming)	250	7
MGC (Vidulin et al., this issue)	1239	20
HGC (Stubbe and Ringlstetter, 2007)	1280	32
KRYS I (Berninger et al., 2008)	5305	70

enable the comparison of different empirical methodologies, to objectively evaluate the performance of different computational models, and, last but not least, to assess the impact of the number of genres, the number of documents, the number of annotators and the criteria of annotation may have on genre findings and on the performance of genre-enabled applications.

3 Genre Classes

The construction of genre benchmarks necessarily involves the task of assigning motivated labels to documents. Although the term "genre can be intuitively understood, huge problems arise when it comes to the identification of which document classes can be considered "genres". For instance, while the Sidney School, centred upon the Systemic Functional Linguistics, e.g., Martin and Rose (2008), focuses more on the role of genre in the linguistic communication system, the North American School e.g., (Swales, 1990; Yates and Orlikowski, 1992) focuses on the genres used in specific communities, e.g., Swales accounts for research and academic genres. Independently from the Systemic School and the North American School, there is an established tradition in German text linguistics of cataloguing genre classes (called *Textsorten*). Görlach, who lists about 2,000 labels (Görlach, 2004), and Adamzik, who collected about 4,500 labels (Adamzik, 1995), belong to this tradition. However, all these nomenclatures or taxonomies seem to be disconnected from the latest trends in automatic genre identification, which is currently handling a proliferation of classes that are not, properly speaking, genres. Some of them have been created in ad-hoc fashion (e.g. *tables* or *lists, person, resources, children, subjective opinion, content delivery*, etc.) because they are assumed to be useful classes when searching the web.

An interesting discussion on this point can be found in (Karlgren, ming), where the author suggests that it not enough to discover new surface features to postulate new genres. Conversely, it is the study of information needs that allow us to detect them, since genres are behavioural categories. On the other hand, (Crowston et al., ming) showed that also user-based genre taxonomies might have their own problems.

4 Research Goals

The major research efforts for the creation of web genre benchmarks are:

1. Propose a characterisation of genre suitable for digital environments and empirical approaches. We pointed out in the Introduction that the lack of a shared and flexible definition of genre is one of the main obstacle to the progress of genre-enabled information systems (see also the discussion in Section 4).
2. Define the criteria for the construction of genre benchmarks and draw up annotation guidelines.
3. Create genre benchmarks in several languages. Genre is cross-cultural concept, so it makes sense to create reference web genre corpora for multiple languages in order to create and evaluate cross-lingual genre-enabled information systems, a promising future direction.
4. Validate the methodology and evaluate the results.

A number of intriguing challenges must be faced in the construction of web genre benchmarks. One challenge is to convey the variety of genre classes that have been used so far in automatic genre classification experiments without cutting out others that can be potentially useful for the information needs of web users (Issue 1). Another challenge is represented by the size of the benchmarks: although designed to be large, benchmark corpora are necessarily limited in size. What is the minimum corpus size (or critical mass) required to test the scalability of genre-enabled applications (Issue 2)? Additionally, we do not know anything about the distribution of genres on the web (Kilgarriff and Grefenstette, 2003), and knowledge about the distribution is essential for machine-learning based systems (Issue 3). Albeit genre colonisation (Beghtol, 2001) is quite extensive on the web, genre classes are social artefacts linked to specific cultures (e.g. it seems that the *obituary* genre is not indigenous to Asian countries), so one must decide on the cross-cultural span of the genre benchmarks (Issue 4). Finally, it is hard to devise benchmarks that are easily updated with the new genres brought about with the advances of web technology (Issue 5). We would like to address these problems as follows:

Issue 1 Diversified genre palettes will be included in the benchmarks, thus allowing a large diversification of genre classes. This variety of genres is useful to test the *portability* of genre classification systems. Additionally, it will enable the study of similarities and differences between genre labels taken from different genre palettes. Possibly, this may lead to the definition of more appropriate and agreed upon genre labels.

Issue 2 Web genre corpora of different sizes will be devised to investigate problems related to scalability and robustness.

Issue 3 We plan a genre-oriented replication of the experiences described in (Thelwall, 2008) to gain new insights into the distribution of genres of the web and reach a better understanding of the dynamics underlying genre use on the web.

Table 2: Examples of mapping from KRYS I to FGC and GCL

FGC/subtype	GCL genre	KRYS I genre
reporting/presentation	Curriculum Vitae, Resume	Resume, CV
reporting/presentation	Encyclopedia	Fact sheet
discussion/academic	Research report, Academic	Academic Monograph
discussion/academic	Research report, Academic	Technical Report
regulations	Contracts, Disclaimer, T & C	Contract

Issue 4 Development of resources for several different languages will allow us to investigate the cultural distance (if any) between the cultures of different countries, and to test cross-lingual genre-enabled information systems.

Issue 5 We plan to monitor genre evolution through a monitor corpus. As the construction of genre monitor corpora is not a trivial issue and has the creation of genre benchmarks as prerequisite, we postpone its creation to future research and projects.

5 The Roadmap

5.1 Short-Term Plan: Mapping existing web genre collections into macro-genres and micro-genres

In the short term, the plan is to capitalise on existing genre-annotated resources. In this "small-scale" work plan we would like to re-utilise the web genre collections listed in Table 1. They are all made of manually annotated English web documents, HTML and PDF. The idea is to provide a stand-off annotation mapping diverse native categories of collections from Table 1 (source) to a set of standardised categories (target) following two genre palettes: the macro-genre of the Functional Genre Classification (FGC), as proposed and motivated in (Sharoff, ming), and the micro-genre of the Genre Classes List (GCL), as presented in (Rehm et al., 2008)[5].

In the end, the majority of documents in each of the six collections will be supplied with their original genre annotation plus two stand-off annotations (see Table 2 for examples). In other words, each existing genre collection will have the original genre annotation decided by its creators (source genre annotation), plus additional genre labels coming from the FGC palette (stand-off macro-genre annotation) and the GCL palette (stand-off micro-genre annotation). This means that we will have about 10,000 webpages with two consistent stand-off annotation schemes.

Any mapping between genre schemes has to accommodate three problematic cases:

[5] Later on, genre labels from other palettes can also be added, e.g., from (Rosso and Haas, ming).

1. a many-to-one mapping, when the target collection does not make finer-grained distinctions made in a source collection;
2. a one-to-many mapping, when the more general class of a source collection can be mapped into more than one genre class in our target palette;
3. a many-to-many mapping, when the two classification schemes are incompatible. For cases of many-to-many mappings between classes we will define additional features needed to achieve unambiguous mapping between individual documents.

Our hypothesis is that in many cases it is possible to design a mapping on the level of classes in each collection, use automatic classification methods for approximate reclassification of more general classes and review their results manually. The proposed harmonisation of genre classes is similar to the comparison of Part-Of-Speech tagsets in the AMALGAM project (Atwell et al., 2000), when a corpus was tagged with 8+ rival tagsets.

In the first stage we have mapped the diverse labels from genre collections of Table 1 to the FGC palette, which includes the following macrogenres and their subtypes:

1. **discussion** – all texts expressing positions and discussing a state of affairs, the three main subtypes are **public** (corresponding to public debates, like blogs or opinionated journalistic texts), **academic** (research papers, books), and **communication** (spontaneous electronic communication, like discussion forums or chat rooms);
2. **reporting** – objective texts reporting on a state of affairs, the two main subtypes are **events** (like newswires and police reports) and **presentation** (like homepages, specifications or CVs);
3. **information** – catalogues, glossaries, sitemaps, other lists of links (mostly containing incomplete or isolated sentences);
4. **instruction** – how-tos, FAQs, tutorials;
5. **propaganda** – adverts, political pamphlets;
6. **recreation** – fiction and popular lore;
7. **regulations** – laws, small print, rules;
8. **unknown** – this was reserved for webpages with little or no natural language, like forms for queries, logins, flash animation, samples of source code, etc.

This palette is compact and coarse-grained, so that it is easier to conflate finer-grained genre classes of each genre collection into coarser-grained functional genres and into their subtypes where possible (see Table 3 for examples)[6].

Nevertheless, the application of the FGC palette to the genre collections listed in Table 1 revealed many cases of ambiguities. Often labels in source collections are ambiguous, and only an investigation of their content can help to determine the target category, e.g., the label `informative` in MGC applies to CVs, descriptions of companies, encyclopedic definitions; similarly, `article` in KI-04 covers research papers, not news articles. In other cases, a single label in a source collection covers webpages of several

[6]The complete stand-off annotation is available from `http://purl.org/net/webgenres`

Table 3: Mapping from different annotation schemes into the FGC palette

FGC macrogenre	FGC subtype	Source genre
		MGC (20 genres)
N/A		adult
discussion	public	blog
recreation		childrens
propaganda	shop	commercial/promotional
discussion	communication	community
unknown		content delivery
recreation		entertainment
unknown		error message
instruction		FAQ
information		gateway
information		index
reporting	presentation	informative
discussion	public	journalistic
N/A		official
N/A		personal
recreation		poetry
recreation		prose fiction
discussion	academic	scientific
propaganda	shop	shopping
N/A		user input
		KI-04 (8 genres)
discussion	academic	article
N/A		discussion
unknown		download
instruction		help
information		linklists
reporting	presentation	portrait-non_priv
reporting	presentation	portrait-priv
propaganda	shop	shop

different genres, so that its target label is not unique, e.g., `adult` in MGC covers lists of links, advertising, forms for accessing websites, legal disclaimers, instructions, etc.

In the second phase of the short-term plan, we would like to utilise experience gained in this process to map to the fine-grained GCL palette from (Rehm et al., 2008). In spite of the more diverse set of genres in GCL, unambiguous mapping is still possible in many cases (see examples in Table 2). However, we envisage much greater need for semi-automatic one-to-many mappings at this stage.

A technical problem inevitable with the unification of diverse genre collections concerns the difference in their storage methods. Some collections include webpages with their respective stylesheets, images and Javascripts, while others include only HTML pages proper. Some collections store files in a hierarchy of directories, while others contain flat lists. We unified the storage methods to the lowest common denominator: HTML pages only in a flat list. For the PDF pages from KRYS-I we created their text versions using `pdftotext`. The stand-off annotation contains ids of HTML files with respective annotation labels.

5.2 Long-Term Plan

Phase I: Discussion, Decisions and Guidelines Building up on the experience accumulated during the short-term plan activities, we will start the long-term plan by building upon the 10,000 web document corpus. This stage will provide a flexible definition of web genre for computational purposes and comprehensive annotation guidelines to reduce the level of ambiguity.

Phase II: Genre Benchmark Construction In this phase, the collection, annotation and storage of the web documents following the criteria defined in Phase I will take place. We anticipate that a number of genre corpora will be built during this phase. While the short term plan focused on English, in this phase we plan the construction of three corpora of web documents in several languages to allow the evaluation of cross-lingual genre-enabled information systems. Provisionally, we call these three corpora: "gold" corpus, "main" corpus and "comprehensive" corpus.

The "gold" corpus for each language will be annotated by several annotators to assist in studies measuring the level of disagreement between annotators, as well as cases of genre hybridism. With this smaller corpus we will also investigate the effect of using radically different genre palettes, i.e. documents will be annotated with codes taken from incompatible sets of genres. It is worth noting that the concept of genre hybridism subsumes several perspectives on text (see Section 6).

Documents in the "main" corpus for each language will be annotated, each following the main annotation schemes resulting from the previous step.

We will also prepare a "comprehensive" corpus (on the order of hundreds of thousands documents), which will be annotated automatically. We will train statistical classification

models on the basis of the "main" corpus, leveraging on semi-supervised machine learning techniques, e.g., boostrapping and active learning, and apply them to the bigger corpus[7].

With the "comprehensive" corpus, we would like to address two research issues:

1. genre hybridism, i.e. several separate genres in a single page, e.g., a newspaper article and a forum discussing it;
2. ambiguity in interpretation, e.g., ambiguity in the genre palette itself, see the description of wikipedia pages in (Rehm et al., 2008).

Importantly, all the corpora will follow a multi-labelling annotation scheme, where web pages are not necessarily (and artificially) restricted to the membership of a single genre. Techniques will be developed to establish sensible labelling thresholds. With the approach proposed above, web pages will be endowed with zero, one or more genre labels, as needed. This will allow future investigators to shed some light on whether the 'nature' of genre and the annotation method affect the performance of genre-enabled applications. Even if the quality of automatic classification in the "comprehensive" corpus is far from perfect, a really big genre-annotated corpus should help researchers estimate the performance of their models on large-scale resources, one of the main holes in current automatic genre research.

Phase III: Creation and Evaluation of Automatic Genre Identification Systems In this phase, the criteria and the experience built up in the previous phases will be used to develop reliable automatic genre classification models. During this phase, new evaluation methods and measure will be proposed to investigate the correlation among different genre granularity and classification schemes. Previous experiments have already shown that computable relations exist between rhetorical genres (like narration or argumentation) and social genres (such as blogs and editorials). For instance, see the two-layer approach proposed by Santini (ming), where these relations have been investigated only on small and heterogeneous genre corpora, which did not allow a robust evaluation of the results. The construction of principled benchmarks will allow us to delve deeply into evaluation techniques and eventually propose new evaluation measures, which more suitably account for classifier performance with difficult classes like genres. It is worth emphasising that multi-labelled genre evaluation is a challenging and very little explored field (an exception is Vidulin et al. in this Issue), and the contribution of this project in this respect will certainly be remarkable. Multi-labelling presents challenges for the current state of machine learning. This is why our project is timely and would complement other initiatives, e.g., see the Workshop on Learning from Multi-Label Data (MLD'09)[8].

[7] A similar approach is used in the ongoing project at the University of Leeds (UK) supported by a Google Research Award for 2009-2010, `http://corpus.leeds.ac.uk/serge/webgenres/google.html`

[8] `http://lpis.csd.auth.gr/workshops/mld09/`

6 Corpus Design Issues

Since the web is a huge reservoir of texts that can be easily mined, we propose building genre benchmarks with freely downloadable web documents. This decision still leaves us with a range of open questions.

Document type Although we are well aware that the web is not limited to HTML pages and PDF files, in this project we will focus on these two document types, leaving the exploration of other types to future research.

Document selection An important open question concerns the criteria for selecting documents. Some researchers have attempted to use equal amount of texts per genre, while others have mined random samples of webpages for a given language or used existing text collections. This project is aimed at producing a set of diversified genre classes, thus resulting in multiple corpora corresponding to multiple benchmarks. In the end, the exact inventory of genres cannot be fixed and the corpus cannot be balanced by this criterion a priori. At the same time, a set of annotated texts from the total set of texts can be selected according to wishes of individual researchers, e.g. the subset chosen by a researcher can contain 200 news items vs. 100 editorials. The second argument in favour of using a random sample from the web for initial annotation is related to the purpose of our benchmarks, which have to reflect the composition of the web to be useful in application domains.

Genre hybridism Genre hybridism is broad term accounting for several phenomena. It has often been pointed out that genres are not discrete systems. A number of genre combinations are possible and common. For example, a mixed genre, like the tragi-comedy, is a genre having its own blending aspects of two or more genres. Multi-genre documents are documents where two or more genres overlap creating a specific and more standardised genre, as in the case of eshops, which are often also search pages. Some genres are intrinsically mixed, such as the newsletter, which contains editorials, reports, interviews, and so on. An additional problem concerns the fuzziness of genre labels because, for example, the same document can be named news bulletin or press release. An account of how difficult can be to build a genre taxonomy is given by Crowston et al. (2009). Hybrid genres abound and are very common in all mass media. In an open environment, such as the web, this phenomenon seems to be pervasive. Generally speaking, the concept of genre hybridism simply helps pin down when a web page contains more than one genre, regardless how these genres relate to each other. The acknowledgement that a web page can be hybrid is important when dealing with automatic genre identification, because traditional single-label classification algorithms are usually confused by hybrid genre conventions. However, at this stage, we have not decided yet whether the produced benchmarks will provide an *ordering* of labels. For example, a page of a newspaper article may contain a discussion forum. Ideally, in this case an ordering could be provided: 1=article and 2=forum. However, this

hierarchical ordering is not always possible, because many web pages often show several unrelated texts, like the ads connected to certain keywords (e.g. see how the application "Google AdWords" works). Other interesting genre information could be provided by the *positioning*, that is, a specific part of the web page is an article and another specific part is a discussion forum. Ordering and positioning information are crucial to evaluate in depth the accuracy of web genre detection tools. However, at present, genre research has not reached the maturity needed to spell out ordering and positioning. We put off these interesting issues to future projects.

Copyright Another crucial question concerns copyright. According to existing copyright law researchers are free to distribute URL links with their descriptions, from which it is possible to recreate a corpus in any necessary format (Sharoff, 2006b). The major problem with this method is that the web changes, some pages get deleted, others updated. An experiment in measuring the decay rate of URLs estimated the half-life of an Internet corpus as about seven years, i.e. the half of the offline webpages of an average collection get changed or deleted in about seven years (Sharoff, 2006a). Storage and redistribution of complete webpages is not traditionally allowed under copyright law. Some Internet corpus projects managed to overcome this constraint by putting sentences in their corpora in random order, for instance, some portions of the Hunglish corpus have been shuffled (Varga et al., 2007). This makes it possible to redistribute the content of webpages with appropriate annotations, but this prevents doing discourse analysis or any other investigation of contexts larger than a sentence. The most suitable solution for development of our reference webgenre corpus is to follow the practice of distribution of other webcorpora, such as deWac (Baroni and Kilgarriff, 2006) or ukWac (Ferraresi et al., 2008), which give the provision for copyright holders of individual webpages to opt out from keeping their pages in the collection. In addition, it is possible to select webpages explicitly marked with permissive licences, such as the GNU Free Documentation Licence or a family of Creative Commons Licences, even though this choice can bias the selection of texts.

Automatic genre identification Traditionally, scholars and researchers studying the genre of documents annotate these documents themselves, i.e. manually. The main drawback with manual annotation is that it is extremely tedious and time-consuming. Consequently, the number of documents manually annotated by genre is often too small to have a full picture of certain phenomena or to carry out any quantitative approach. Additionally, now with the web and with the wealth of freely available documents, the 'manual annotation pace' is certainly a huge limitation for genre research. The second drawback is that since manual annotation is a mentally demanding activity, tiredness or distraction causes errors and idiosyncrasies. Ideally, as machines do not get tired, they should provide genre analysts with larger quantity of consistently genre-annotated documents. In brief, annotating documents by genre is not always an easy task: it takes time, it is not always intuitive and it is prone to errors, because human annotators get

easily tired or confused. For this reason, automatic genre classifiers would be a great advantage in building web genre benchmarks.

7 Significance of the Research and Conclusion

This project will provide the community of genre scholars and practitioners with a number of theoretical contributions, and several valuable resources.

From a theoretical point of view, this project will enrich genre studies and genre research with a characterisation of the concept of genre tailored for digital environments. It will also produce a set of re-usable criteria for the construction of web genre benchmarks and annotation guidelines, so that computational experiments can be carried out with a large number of diverse web documents. Additionally, it will provide a comparative assessment of a range of existing genre annotation schemes with a mapping between these onto a neutral palette. We conjecture that significant insights will be yielded by the experiments tested on such a resource.

Last but not least, it will provide long-lasting web document collections, namely a number of web genre benchmarks in several languages, which can be updated, monitored and enlarged in future. Importantly, in this article we describe work in progress and our plans for future development. Since it is sometimes difficult to anticipate the difficulties that will arise when developing a large resource, we present our ideas, our current views on genre issues and our first results with the aim of stimulating a proactive discussion, so that the stakeholders, i.e. researchers who will ultimately benefit from the resource, can contribute to its design.

References

Adamzik, K. (1995). *Textsorten – Texttypologie. Eine kommentierte Bibliographie*. Nodus, Münster.

Atwell, E., Demetriou, G., Hughes, J., Schriffin, A., Souter, C., and Wilcock, S. (2000). A comparative evaluation of modern english corpus grammatical annotation schemes. *ICAME Journal*, 24:7–23.

Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. In *Companion Volume to Proc. of the European Association of Computational Linguistics*, pages 87–90, Trento.

Bateman, J. (2008). *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*. Palgrave Macmillan.

Beghtol, C. (2001). The concept of genre and its characteristic. *Bulletin of ASIST*, 27(2):17–19.

Berninger, V., Kim, Y., and Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. In *Proceedings of the Corpus Profiling Workshop*, London.

Braslavski, P. (Forthcoming). Marrying relevance and genre rankings: an exploratory study. In Mehler et al. (ming).

Crowston, K., Kwaśnik, B., and Rubleske, J. (Forthcoming). Problems in the use-centered development of a taxonomy of web genres. In Mehler et al. (ming).

Dimitrova, M. and Kushmerick, N. (2003). Dimensions of web genre. In *World Wide Web Conference WWW2003, Budapest, Hungary*.

Erickson, T. (1999). Rhyme and punishment: the creation and enforcement of conventionsin an on-line participatory limerick genre. In *Proc. 32nd Annual Hawaii International Conference on System Sciences*.

Ferraresi, A., Zanchetta, E., Bernardini, S., and Baroni, M. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *The 4th Web as Corpus Workshop: Can we beat Google? (at LREC 2008)*, Marrakech.

Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11).

Finn, A., Kushmerick, N., and Smyth, B. (2002). Genre classification and domain transfer for information filtering. In *Proc. European Colloquium on Information Retrieval Research*, Glasgow.

Freund, L. (2008). *Exploiting task-document relationships to support information retrieval in the workplace*. PhD thesis, University of Toronto.

Görlach, M. (2004). *Text types and the history of English*. Walter de Gruyter.

Gupta, S., Becker, H., Kaiser, G., and Stolfo, S. (2006). Verifying genre-based clustering approach to content extraction. In *Proceedings of the 15th international conference on World Wide Web*, pages 875–876. ACM.

Heyd, T. (2008). *Email hoaxes: form, function, genre ecology*. Benjamins.

Kanaris, I. and Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management*, 45:499–512.

Karlgren, J. (2005). The whys and wherefores for studying textual genre computationally. In *Proc. AAAI Fall Symposium on Style and Meaning in Language, Art and Music*, Arlington, USA.

Karlgren, J. (Forthcoming). Conventions and mutual expectations — understanding sources for web genres. In Mehler et al. (ming).

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., and Wolkert, N. (1998). Iterative information retrieval using fast clustering and usage-specific genres. In *Eight DELOS workshop on User Interfaces in Digital Libraries*, pages 85–92.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of the 15th. International Conference on Computational Linguistics (*Coling 94*)*, pages 1071 – 1075, Kyoto, Japan.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th ACL/8th EACL*, pages 32–38.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue of the web as corpus. *Computational Linguistics*, 29(3):333–347.

Kim, Y. and Ross, S. (Forthcoming). Formulating representative features with respect to genre classification. In Mehler et al. (ming).

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Martin, J. and Rose, D. (2008). *Genre Relations: mapping culture*. Equinox Pub.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. (2001). Named entity recognition from diverse text types. In *Proc. Recent Advances in Natural Language Processing*, pages 257–274.

Mehler, A., Gleim, R., and Wegner, A. (2007). Structural uncertainty of hypertext types. an empirical study. In *Proc. Towards Genre-Enabled Search Engines: The Impact of NLP. RANLP-07*.

Mehler, A., Sharoff, S., and Santini, M., editors (Forthcoming). *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Meyer zu Eissen, S. and Stein, B. (2004). Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*, Ulm, Germany.

Rauber, A. and Müller-Kögler, A. (2001). Integrating automatic genre analysis into digital libraries. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10.

Rehm, G. (2002). Towards automatic web genre identification – a corpus-based approach in the domain of academia by example of the academic's personal homepage. In *Proc. of the Hawaii Internat. Conf. on System Sciences*.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech.

Rosso, M. A. and Haas, S. W. (Forthcoming). Identification of web genres by user warrant. In Mehler et al. (ming).

Santini, M. (Forthcoming). Cross-testing a genre classification model for the web. In Mehler et al. (ming).

Sharoff, S. (2006a). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. http://wackybook.sslmit.unibo.it.

Sharoff, S. (2006b). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Sharoff, S. (Forthcoming). In the garden and in the jungle. Comparing genres in the BNC and Internet. In Mehler et al. (ming).

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Stein, B., Meyer zu Eissen, S., and Lipka, N. (Forthcoming). Web genre analysis: Use cases, retrieval models, and implementation issues. In Mehler et al. (ming).

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. In *Abstract Proceedings of the Colloqium "Towards a Reference Corpus of Web Genres*.

Swales, J. (1990). *Genre Analysis. English in academic and research settings*. Cambridge University Press, Cambridge.

Symonenko, S. (2007). Recognizing genre-like regularities in website content structure. In *Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*.

Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11):1702–1710.

Toms, E. and Campbell, D. (1999). Genre as interface metaphor: exploiting form and function indigital environments. In *Proc. 32nd Annual Hawaii International Conference on System Sciences*.

Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., and Tron, V. (2007). Parallel corpora for medium density languages. In N. Nicolov, K. Bontcheva, G. A. and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05*, pages 247–258. Benjamins.

Yates, J. and Orlikowski, W. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of management review*, pages 299–326.

List of Contributors

Stephen W. Briner
Department of Psychology
DePaul University
Chicago, IL 60614, USA
sbriner@depaul.edu

Philip M. McCarthy
Department of English
The University of Memphis
Memphis, TN 38152, USA
pmmccrth@memphis.edu

Malcolm Clark
The Robert Gordon University
School of Computing
Schoolhill, Aberdeen, AB10 1FR, Scotland, UK
St Andrews Street Building C43
m.clark1@rgu.ac.uk

Michal Cutler
Department of Computer Science
SUNY at Binghamton
Binghamton, NY 13902-6000, USA
cutler@binghamton.edu

Matjaž Gams
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
matjaz.gams@ijs.si

Arthur C. Graesser
Department of Psychology
The University of Memphis
Memphis, TN 38152, USA
graesser@memphis.edu

Patrik O'Brian Holt
The Robert Gordon University
School of Computing
Schoolhill, Aberdeen, AB10 1FR, Scotland, UK
St Andrews Street Building C43
ph@comp.rgu.ac.uk

Chaker Jebari
King Saud University
College of Computer and Information Sciences, KSA
jebarichaker@yahoo.fr

Ryan Levering
Department of Computer Science
SUNY at Binghamton
Binghamton, NY 13902-6000, USA
ryan.levering@binghamton.edu

Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
mitja.lustrek@ijs.si

Danielle S. McNamara
Department of Psychology
The University of Memphis
Memphis, TN 38152, USA
dsmcnamr@memphis.edu

Alexander Mehler
Faculty of Technology
Bielefeld University
33615 Bielefeld, Germany
alexander.mehler@uni-bielefeld.de

John C. Myers
Department of Psychology
The University of Memphis
Memphis, TN 38152, USA
jcmyers@memphis.edu

Silvia Petri
Dipartimento di Studi italianistici
Universita di Pisa
Via del Collegio Ricci 10
Pisa, Italy
silvia.li@inwind.it

Georg Rehm
vionto GmbH
Karl-Marx-Allee 90a
10243 Berlin, Germany
georg.rehm@vionto.com

Ian Ruthven
Computer and Information Sciences
University of Strathclyde
16 Richmond Street
Glasgow G1 1XQ, Scotland, UK
ian.ruthven@cis.strath.ac.uk

Marina Santini
Humanities Advanced Technology
and Information Institute (HATII)
University of Glasgow
Glasgow, Scotland, UK
marinasantini.ms@gmail.com

Serge Sharoff
Centre for Translation Studies
University of Leeds
Leeds, UK
s.sharoff@leeds.ac.uk

Mirko Tavosanis
Dipartimento di Studi italianistici
Universita di Pisa
Via del Collegio Ricci 10
Pisa, Italy
tavosanis@ital.unipi.it

Vedrana Vidulin
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
vedrana.vidulin@ijs.si